Molecular Genetics and Genomics (2019) 294:1343–1357
https://doi.org/10.1007/s00438-019-01584-6

ORIGINAL ARTICLE

Population genetics, diversity and forensic characteristics of Tai–Kadai-speaking Bouyei revealed by insertion/deletions markers

Guanglin He[1,2] · Zheng Ren[3] · Jianxin Guo[2] · Fan Zhang[3] · Xing Zou[1] · Hongling Zhang[3] · Qiyan Wang[3] · Jingyan Ji[3] · Meiqing Yang[3] · Ziqian Zhang[2] · Jing Zhang[2] · Yilizhati Nabijiang[2] · Jiang Huang[3] · Chuan-Chao Wang[2]

Received: 25 December 2018 / Accepted: 30 May 2019 / Published online: 13 June 2019

Abstract

China, inhabited by over 1.3 billion people and known for its genetic, cultural and linguistic diversity, is considered to be indispensable for understanding the association between language families and genetic diversity. In order to get a better understanding of the genetic diversity and forensic characteristics of Tai–Kadai-speaking populations in Southwest China, we genotyped 30 insertion/deletion (InDel) markers and amelogenin in 205 individuals from Tai–Kadai-speaking Bouyei people using the Qiagen Investigator DIPplex amplification kit. We carried out a comprehensive population genetic relationship investigation among 14,303 individuals from 84 worldwide populations based on allele frequency correlation and 4907 genotypes of 30 InDels from 36 populations distributed in all continental or major subregions and seven linguistic phyla in China. Forensic parameters observed show highly polymorphic and informative features for Asians, although the DIPplex kit was developed focusing on Europeans, and indicate that this amplification system is appropriate to forensic personal identification and parentage testing. Patterns of InDel variations revealed by principal components analysis, multidimensional scaling plots, phylogenetic relationship exploration, model-based clustering as well as four pairwise genetic distances (Fst, Nei, Cavalli-Sforza and Reynolds) demonstrate significant genetic differentiation at the continental scale and genetic uniformity in Asia except for Tibeto-Burman and Turkic-speaking populations. Additionally, Tai–Kadai speakers, including Bouyei, Zhuang and Dong, share more genetic ancestry components than with other language speakers, and in general they are genetically very similar to Hmong–Mien-speaking populations. The dataset of Bouyei people generated in the present study is valuable for forensic identification and parentage tests in China.

Keywords InDels · Population structure · Tai–Kadai · Forensic genetics · Population genetics · Linguistic family

Introduction

The out of Africa and peopling other continents of anatomically modern human have been subsequently evidenced by the genetic (maternal, paternal and biparental autosomal variations), archaeogenetic and linguistic advances (Nielsen et al. 2017). Understanding the patterns of variation and genetic structure of modern human globally is a complex task which needs data from multiple linguistically, ethnically, geographically diverse populations. China, located in East Asia, was first inhabited by hunter-gatherers from approximately 50 thousand years ago (Kya) and then occupied by farmers via southward and northward expansions from the Yangtze River Basin and the Upper and Middle of Yellow River Basin in the late Neolithic and Bronze age (Su et al. 1999; Barton et al. 2009; Yang et al. 2012). Modern Chinese populations consist of 55 ethnic groups

Communicated by Stefan Hohmann.

Guanglin He and Zheng Ren contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00438-019-01584-6) contains supplementary material, which is available to authorized users.

✉ Jiang Huang
mmm_hj@126.com

✉ Chuan-Chao Wang
wang@xmu.edu.cn

[1] Institute of Forensic Medicine, West China School of Basic Science and Forensic Medicine, Sichuan University, Chengdu 610041, China

[2] Department of Anthropology and Ethnology, Institute of Anthropology, Xiamen University, Xiamen 361005, China

[3] Department of Forensic Medicine, Guizhou Medical University, Guiyang 550004, Guizhou, China

and the world's largest Han Chinese, which have a population of over 1.4 billion. The language landscape of China is dominated by at least ten language families containing 292 different languages, including Sino–Tibetan, Tai–Kadai, Hmong–Mien, Austroasiatic (Wa), Indo-European (Tajiks), Austronesian (Taiwanese), Tungusic, Turkic, Mongolic and Koreanic. The ancient DNA analysis of an upper Paleolithic human sample from Tianyuan Cave (40 Kya) and human genetic diversity in prehistoric East Asia indicated that the Paleolithic Chinese shared more alleles with ancient and modern Asians and there were a complex migration and subdivision of early Asian populations (Yang et al. 2017). The two early Neolithic hunter-gathering East Asian individuals (dated to 7.7 Kya) in Russian Far East were genetically close to modern Amur Basin Tungusic-speaking populations, indicating the genetic continuity in Asia, which is not like complex population turnover reported in Britain and remote Oceania (Siska et al. 2017; Lipson et al. 2018b; Olalde et al. 2018). The reconstruction of population prehistory in Southeast Asia based on two Hòabìnhian hunters-gatherers, 41 rice and millet farmers and one Japanese Jōmon revealed that the complex genetic makeup of Southeast Asians were due to multiple incoming waves of East Asians (Lipson et al. 2018a; McColl et al. 2018). Recent genetic studies have also shown that the mixed ancestry landscape of present-day Chinese was mediated by the major episodes of gene flow from Southeast Asia, South Asia, Western Siberia and Eastern Siberia (Lu et al. 2016; Feng et al. 2017). Other inconsistent evidence from historical linguistics, archeologists and geneticists focused on the origin and spread of Chinese languages, dynamics of material culture, population structure and genetic history of ethnolinguistically and geographically diverse Chinese populations have promoted China as a research hotspot.

Insertion/deletion polymorphisms (InDels, DIPs) harbor the features of smaller amplicons, lower mutation rates and higher performance like single nucleotide polymorphisms (SNPs). InDels also possess the characteristics of fragment length polymorphisms suitable for separation in the capillary electrophoresis (CE) technology like forensic gold standard short tandem repeats (STRs). Therefore, InDels have attracted and gained increasing attention in the population genetic and forensic community (Zhu et al. 2018). Human diallelic InDel polymorphisms were first identified and reported by Weber in 2002. And then they found that these binary markers consist of approximately 8% of human genome variations (Weber et al. 2002). Mills et al. followingly conducted a comprehensive mapping via over 4.5 million InDels, which facilitated the genotyping technologies and applications in forensic, medical and population genetics (Mills et al. 2006). Due to the rapid advances in the next generation sequencing, international incorporation achievements in the 1000 genomes project (Sudmant et al. 2015), Simons Genome Diversity Project (SGDP) (Mallick et al. 2016) and Estonian Biocentre Human Genome Diversity Panel (EGDP) (Pagani et al. 2016) have carefully identified and prioritized the distribution patterns of genetic variation of InDels in different worldwide populations. The Investigator DIPplex® kit (Qiagen) including 30 diallelic InDels and Amelogenin was launched for forensic applications in Europeans (Fondevila et al. 2012). Subsequent population genetic studies and forensic performance investigations have been conducted in the Africans, Europeans, Asians, north and south Americans (Akhteruzzaman et al. 2013; Martinez-Cortes et al. 2016; Shen et al. 2016; Xie et al. 2018).

Although a large number of forensic reference database and forensic characteristics of ethnically, linguistically and geographically diverse populations had been generated using the Investigator DIPplex system, the genetic polymorphisms and forensic efficiency of 30 InDels in Tai–Kadai-speaking Bouyei population, the 11th largest ethnic group in China with a population over 2.9 million, remain uncharacterized. Besides, the association between the genetic diversity and language family boundary in this genetically, linguistically and culturally diverse region of Southwest China is still unclear. Thus, we genotyped DIPplex InDel polymorphisms in 205 Guizhou Bouyei individuals and evaluated the forensic efficiency in the present work. Additionally, we first combined our genotype data with other 4702 genotypes of the 30 InDel markers from 36 populations. We then merged our allele frequency dataset with previously investigated data from 84 worldwide populations consisting of 14,303 individuals. We aim to dissect the genetic structure of Bouyei speakers, explore the genetic relationships between Bouyei population and reference groups in the context of worldwide and Asian populations, and finally analyze the genetic similarities and dissimilarities in populations from different language families and illuminate how Tai–Kadai Bouyei has influenced by or been influenced the genetic architecture of neighboring populations.

Materials and methods

Sample collection and DNA preparation

A total of 205 peripheral anticoagulant blood samples (119 males and 86 females) were collected from unrelated healthy voluntary donors residing in Guizhou Province, Southwest China (Fig. 1), with the approval of the Ethics Committee of the Guizhou Medical University. Genomic DNA was isolated using a salting-out procedure and subsequently quantified applying a Nanodrop ND-2000 spectrophotometer (Thermo Fisher Scientific, Wilmington, USA).

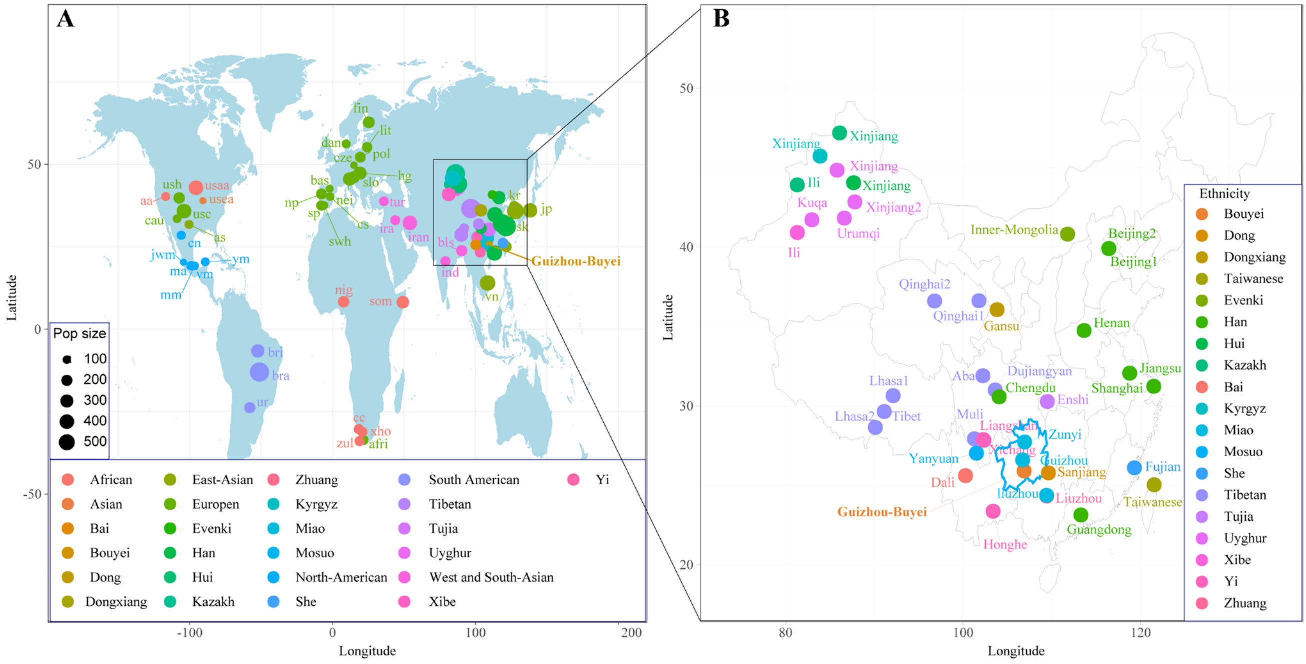

Fig. 1 Geographical distribution of Guizhou Bouyei and other 83 worldwide reference populations. Color of circle in the worldwide map (**a**) and China map (**b**) indicates the ethnic origin or geographical origin of 84 included populations. Size of circle (**a**) indicates the population size. Guizhou Province is located in the Southwest China and corresponding geographical is marked with light blue color (color figure online)

Amplification, genotyping and quality control

Thirty-one forensic related markers were simultaneously amplified on a GeneAmp PCR System 9700 (Thermo Fisher Scientific, Wilmington, USA) using the Qiagen Investigator DIPplex amplification kit on the basis of the manufacturers' specifications with a modified amplification volume of 10 ul. Amplified fragments were isolated using the capillary electrophoresis in an Applied Biosystems 3500 Genetic Analyzer (Thermo Fisher Scientific, Wilmington, USA) following the instructions. InDels nomenclatures and profiling were carried out using the GeneMapper ID-X software (Thermo Fisher Scientific, Wilmington, USA). The positive control is employed as DNA 007 (Thermo Fisher Scientific, Wilmington, USA) and negative control of ddH$_2$O for each batch of amplification and genotyping. We strictly followed the recommendations of the DNA Commission of the International Society of Forensic Genetics (ISFG) on the analysis of forensic markers and internal quality control requirements.

Datasets

We prepared the following two datasets: Dataset I based on raw genotype data and Dataset II based on the allele frequency distribution of 30 InDels, representing populations from different continental divisions or language families. Dataset I contains 4907 samples, including 205 reported here for the first time and 4702 samples extracted from previously published work. Dataset II comprises 14,303 worldwide individuals from 84 populations. Dataset I includes 25 Chinese populations and Dataset II contains 42 Chinese populations, which are widely distributed in geographically and linguistically different administrative regions. Those reference populations belong to Tibeto-Burman, Sinitic, Hmong–Mien, Austronesian, Tai–Kadai, Austroasiatic, Turkic, Tungusic and Mongolic language families, and were used to investigate the relation between language classifications and genetic differences. The geographical information and sample sizes are presented in Fig. 1.

Statistical analysis

Genetic diversity parameters, allele frequencies of 30 markers and corresponding forensic genetic parameters [match probability (MP), power of exclusion (PE), discrimination power (DP), polymorphic information content (PIC), and Typical Paternity Index (TPI)] were calculated using a convenient online tool for STR analysis for Forensics (STRAF) (Gouy and Ziegler 2017). Hardy–Weinberg equilibrium (HWE) and linkage disequilibrium (LD) as well as observed and expected heterozygosities were estimated using Arlequin v.3.5 (Excoffier and Lischer 2010). Population genetic relationship investigations via raw genotype data were conducted using the STRAF software, including pairwise Fst distance developed by Weir and Cockerham and principal component analysis (PCA) (Gouy and Zieger 2017). Three widely employed pairwise Nei, Cavalli-Sforza and Reynolds genetic distances (Nei 1978; Reynolds et al. 1983; Kalinowski 2002) were evaluated using the Phylogeny Inference Packages (gendist package) implemented in PHYLIP version 3.5 on the basis of the allele frequency distribution of 30 InDels among 84 worldwide populations (Excoffier and Lischer 2010). The genetic affinity between Guizhou Bouyei and reference populations was also explored via PCA based on the allele frequency correlation using the Multivariate Statistical Package (MVSP) version 3.22 software (Kovach 2007). Multidimensional scaling plots (MDS) were conducted on the IBM SPSS Statistics 21 (Hansen 2005) and phylogenetic relationship reconstruction was carried out on the Molecular Evolutionary Genetics Analysis Version 7.0 (Mega 7.0) (Kumar et al. 2016) on the basis of four different pairwise genetic distances to further dissect the genetic similarity and differences. Finally, we used the software Structure version 2.3.4.21 (Evanno et al. 2005) to assess the apportionment of genetic ancestry among 4907 individuals from 36 populations. We carried out genetic structure dissection using the Structure software with 100,000 burn-in steps and 100,000 repetitions for the Markov chain Monte Carlo (MCMC). We used K values ranging from 2 to 8 with the ‘independent allele frequencies’ and ‘LOCPRIOR’ models and used the online tool of Structure Harvester to choose the optimized K (Earl and vonHoldt 2011). We used a cluster matching and permutation program (CLUMPP version 1.1.222) (Jakobsson and Rosenberg 2007) as well as the Distruct version 1.1.23 (Rosenberg 2004) to visualize the cluster membership coefficients (*Q*).

Results

Genetic diversity, allele frequency divergence and forensic efficiency

The 205 new individual profiles of 30 InDels from Tai–Kadai-speaking Guizhou Bouyei were obtained using the Qiagen Investigator DIPplex amplification kit and submitted in Supplementary Table S1. After applying the Bonferroni correction of multiple tests (Table 1 and Supplementary Tables S2), no deviations are observed in the HWE (0.05/30 = 0.0017) and LD (0.05/435 = 0.0001). Our results indicate that all 30 InDels in this HWE population are independently inherited. Thus, we can confidently and effectively estimate both the forensic characteristics of the single locus and the combined forensic efficiency indexes in our following analyses. The allelic frequency distribution and corresponding forensic parameters of 30 InDels are submitted in Table 1. Allele frequencies of 30 markers range from 0.0854 for HLD39 insertion allele to 0.9073 for HLD99 insertion allele. The most informative locus is HLD136 with the PIC of 0.3750, while the lowest polymorphic locus is HLD39 with the PIC of 0.1440. The PM, PD and PE values span from 0.3517 (HLD136) to 0.7324 (HLD39), 0.2676 (HLD39) to 0.6483 (HLD136) and 0.0158 (HLD39) to 0.2466 (HLD92), respectively. TPI values vary from 0.5824 at the locus of HLD39 to 1.1389 at the locus of HLD92. The observed and expected heterozygosities span from 0.1415 to 0.5610, and from 0.1565 to 0.5012, respectively.

Genetic markers with different allele frequency among diverse populations can be used as ancestry informative markers (AIM) to determine the individual’s ancestry in a forensic case (Phillips 2015). A considerable number of ancestry inference panels based on the ancestry informative single nucleotide polymorphisms (AISNPs), ancestry informative insertion/deletions (AIDIPs) or ancestry informative multi-InDels have been developed for forensic applications (Phillips 2015; Inacio et al. 2017). To explore the potential for forensic ancestry inference of 30 InDels in this studied panel, we analyzed the insertion allele frequency divergence among 84 worldwide populations (Fig. S1) and calculated unbiased Fst for each InDel locus (Table 1). Cluster I (HLD39, HLD111 and HLD122, Fst > 0.0941) and Cluster IV (HLD125, HLD58, HLD83, HLD133 and HLD97, Fst > 0.3012) show significant allele frequency differences between North Americans and others, and can be used as American-specific AIMs for distinguishing them. Cluster II (HLD114, HLD128, HLD48 and HLD131, Fst > 0.0218) and Cluster V (HLD136, HLD70, HLD56, HLD88 and HLD93, Fst > 0.0128) with obvious frequency differences among different continental populations can be used for separating African and other groups. Cluster VII (HLD118, Fst = 0.1877) and Cluster VIII (HLD67, HLD84, HLD99, HLD64, HLD81, Fst > 0.0354) show significant frequency divergence among Asians and others, which can be used as Asian-specific AIDIPs for ancestry inference between Asians and other groups. The remaining two clusters, Cluster III (HLD77, HLD6, HLD92) and Cluster VI (HLD40, HLD12 and HLD45) display highly informative and polymorphic features with the balanced allele frequency distribution among worldwide populations, which can be used as “identify informative InDels (IIDIPs)”. We also identify stable and balanced frequency distribution of all 30 InDel markers in Europeans, Central Asians and Siberians, which are consistent with the initial purpose of developing this kit for individual identification and parentage testing focusing on Europeans (Fondevila et al. 2012).

Table 1 The allele frequency distribution and corresponding statistical parameters of forensic interest of 30 InDels in the Bouyei population residing in Guizhou Province, Southwest China

Locus	PIC	PM	PD	PE	TPI	Insertion	Deletion	Ho	He	*p*	Fst
HLD101	0.3696	0.3624	0.6376	0.1429	0.8991	0.4268	0.5732	0.4439	0.4905	0.2019	0.0057
HLD111	0.1824	0.6654	0.3346	0.0247	0.6101	0.1146	0.8854	0.1805	0.2035	0.1534	0.2560
HLD114	0.3169	0.4461	0.5539	0.1169	0.8402	0.2707	0.7293	0.4049	0.3958	0.8595	0.0416
HLD118	0.1540	0.7123	0.2877	0.0189	0.5925	0.9073	0.0927	0.1561	0.1686	0.3899	0.1877
HLD122	0.1883	0.6495	0.3505	0.0298	0.6250	0.1195	0.8805	0.2000	0.2110	0.4999	0.0941
HLD124	0.3731	0.3801	0.6199	0.1896	1.0049	0.5439	0.4561	0.5024	0.4974	0.8884	0.0099
HLD125	0.3648	0.3987	0.6013	0.1940	1.0149	0.4000	0.6000	0.5073	0.4812	0.4707	0.0650
HLD128	0.3271	0.4216	0.5784	0.0845	0.7649	0.2902	0.7098	0.3463	0.4130	0.0267	0.0461
HLD131	0.3403	0.4025	0.5975	0.1053	0.8135	0.3195	0.6805	0.3854	0.4359	0.1079	0.0243
HLD133	0.3577	0.3867	0.6133	0.1465	0.9071	0.3707	0.6293	0.4488	0.4677	0.6538	0.0302
HLD136	0.3750	0.3517	0.6483	0.1429	0.8991	0.4951	0.5049	0.4439	0.5012	0.1282	0.0145
HLD39	0.1440	0.7324	0.2676	0.0158	0.5824	0.0854	0.9146	0.1415	0.1565	0.1662	0.1351
HLD40	0.3648	0.3847	0.6153	0.1690	0.9579	0.6000	0.4000	0.4781	0.4812	1.0000	0.0339
HLD45	0.3667	0.3854	0.6146	0.1771	0.9762	0.5902	0.4098	0.4878	0.4849	1.0000	0.0419
HLD48	0.3706	0.4040	0.5960	0.2216	1.0789	0.4341	0.5659	0.5366	0.4925	0.2041	0.0218
HLD56	0.3737	0.3922	0.6078	0.2121	1.0567	0.4634	0.5366	0.5268	0.4985	0.4884	0.0329
HLD58	0.3744	0.3681	0.6319	0.1730	0.9670	0.5244	0.4756	0.4829	0.5000	0.6735	0.0764
HLD6	0.3676	0.3617	0.6383	0.1327	0.8761	0.4146	0.5854	0.4293	0.4866	0.1144	0.0091
HLD64	0.1998	0.6311	0.3689	0.0325	0.6327	0.8707	0.1293	0.2098	0.2257	0.3460	0.2155
HLD67	0.2839	0.4914	0.5086	0.0687	0.7270	0.7805	0.2195	0.3122	0.3435	0.2180	0.0378
HLD70	0.3684	0.3821	0.6179	0.1771	0.9762	0.5805	0.4195	0.4878	0.4882	1.0000	0.0320
HLD77	0.3469	0.4141	0.5859	0.1612	0.9404	0.3366	0.6634	0.4683	0.4477	0.5339	0.0094
HLD81	0.2910	0.4799	0.5201	0.0774	0.7482	0.7707	0.2293	0.3317	0.3543	0.4291	0.1694
HLD83	0.3577	0.3982	0.6018	0.1690	0.9579	0.3707	0.6293	0.4781	0.4677	0.7646	0.0777
HLD84	0.2766	0.5032	0.4968	0.0645	0.7168	0.7902	0.2098	0.3024	0.3323	0.2113	0.0354
HLD88	0.3726	0.3635	0.6365	0.1574	0.9318	0.5488	0.4512	0.4634	0.4965	0.3995	0.0156
HLD92	0.3750	0.4111	0.5889	0.2466	1.1389	0.4951	0.5049	0.5610	0.5012	0.0932	0.0141
HLD93	0.3676	0.3884	0.6116	0.1854	0.9951	0.5854	0.4146	0.4976	0.4866	0.7779	0.0128
HLD97	0.3590	0.3781	0.6219	0.1327	0.8761	0.3756	0.6244	0.4293	0.4702	0.2317	0.0392
HLD99	0.1540	0.7025	0.2975	0.0235	0.6065	0.9073	0.0927	0.1756	0.1686	1	0.0784

PIC polymorphism information content, *PM* random matching probability, *PD* power of discrimination, *PE* power of exclusion, *TPI* Typical Paternity Index, *Ho* observed heterozygosity, *He* expected heterozygosity, *p p* value in the Hardy–Weinberg

Genetic structure and population relationships in the context of Eurasians and Americans

We first contextualized the new data with 4702 genotypes from published sources. We used principal components analysis, multidimensional scaling plots, heatmap plot and model-based structure dissection to assess the extent of genetic homogeneity and heterogeneity among Tai–Kadai-speaking populations with diverse groups from America, Europe and Asia, and the amount of ancestry sharing among and within adjacent populations. In PCA, the top three components, extracting a total of 18.86% variances, reveal three indistinct clusters, encompassing continental geographically adjacent populations: European and Turkic-speaking groups, Americans, and Asians. American populations are mainly localized in the upper right position and Asians are grouped in the upper left position. The European and Asian admixed Turkic-speaking populations are intermediate between them. The Tai–Kadai-speaking Bouyei population is placed within the Asian cluster (Fig. 2a, b), especially falling together with the Zhuang and Dong populations. A similar pattern of genetic relationships is observed in pairwise Fst genetic distance result (Fig. 2c). The largest genetic distances are observed between American populations, followed by Europeans and Turkic-speaking populations. Focusing on Asian populations, we identify consistently lower Fst values between Bouyei and other Tai–Kadai speakers (Guangxi Zhuang 0.0018, Guangxi Dong 0.0031), following Hmong–Mien-speaking She (0.0034), compared to other Chinese populations (Table S3).

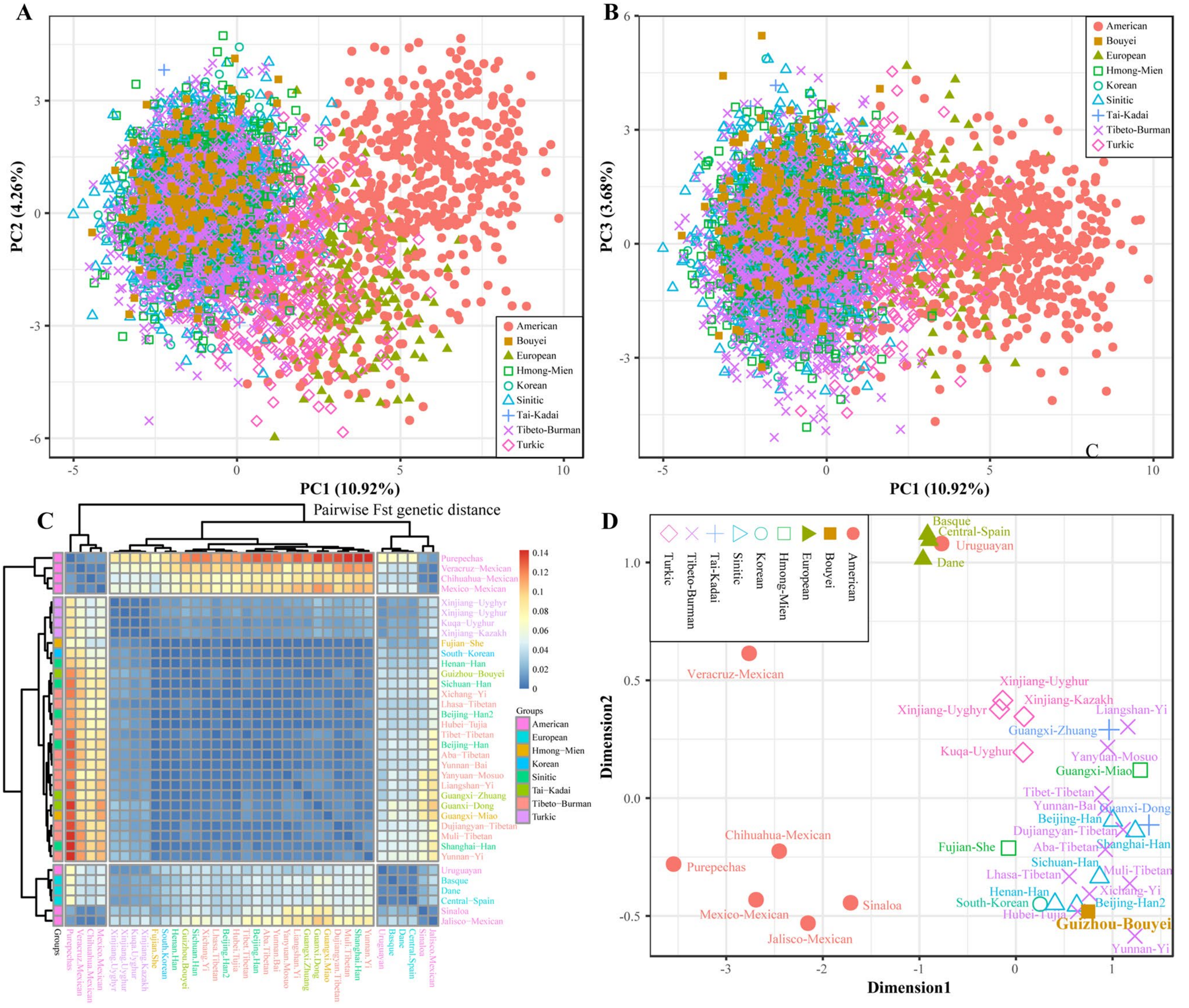

Fig. 2 Genetic affinity between Guizhou Bouyei and other 35 reference populations based on raw genotype data. **a**, **b** Principal component analyses (PCA) among 4907 individuals based on the top three components; **c** the heatmap of pairwise Fst genetic distances among 36 populations; **b** multidimensional scaling plots on the basis of Fst distance matrix

To clearly visualize the genetic structure at the population level, we followingly carried out the MDS and phylogenetic relationship reconstruction in the context of Fst genetic distance matrix. A two-dimensional plot of population relationship distribution is presented in Fig. 2d. American populations are localized at the left lower corner with the exception of Veracruz Mexican and Uruguayan. Three European populations clustered with Uruguayan are placed in the upper middle position, while East Asian groups are placed in the right lower corner. Four Turkic-speaking Kazakh and Uyghurs are placed between Americans and East Asians and they fall closely with Tibetan–Burman-speaking populations. We identify two main branches in the reconstructed neighbor-joining tree: American and European cluster, and Asian cluster. As shown in Fig. 3a, strong correlations between genetic affinity and linguistic similarity are identified within the Turkic-, Tibetan–Burman-speaking populations except for Hubei Tujia and Yunnan Yi. A mosaic relationship is detected between linguistic affiliation and genetic affinity in Sinitic-, Hmong–Mien- and Tai–Kadai-speaking populations. Guizhou Bouyei population first clusters with Guangxi Dong and then consolidates with Guangxi Zhuang.

Additionally, we conducted a structure-like clustering analysis to estimate ancestry proportions among Bouyei and reference populations via InDels raw data. At $k=2$, two major ancestry components have been indentified: American-dominant component (shown in DeepSkyBlue) and East Asian-dominant component (shown in CadetBlue in $k=2$, Fig. 3b).

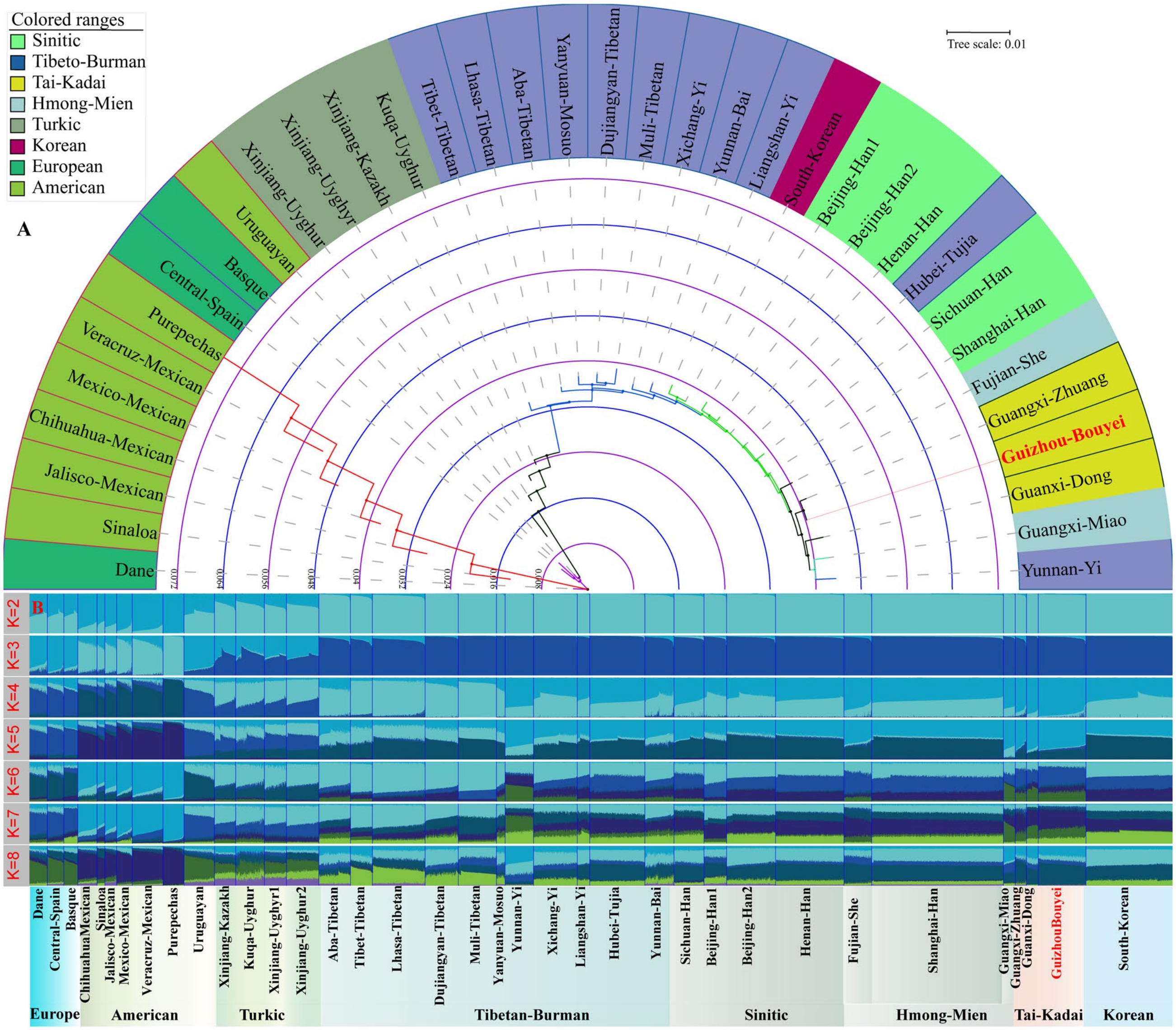

Fig. 3 Phylogenetic relationship reconstruction and structure results. **a** Phylogenetic tree reconstructed based on the neighbor-joining algorithm. **b** Ancestry component proportion in 36 populations with k ranging from 2 to 8

The American-dominant component also has a comparable proportion presented in the Europeans and Turkic-speaking Asians, but is almost absent in other East Asians. At $k=3$, European populations are obviously separated by the European-dominant ancestry component (labeled by DeepSkyBlue in the $k=3$ panel), which shows an equal proportion in Uruguayan population and the relative proportion in American populations. We can also observe approximately 50% of European ancestry and 50% of East Asian ancestry in Turkic-speaking populations, which are in accordance with the Uyghur admixture history revealed by the genetic variations in the chromosome 21, whole-genome high-density SNPs and ancestry informative SNPs (Xu et al. 2008; Xu and Jin 2008; He et al. 2018a). At $k=4$, which is the best appropriate predefined K (Fig. S2), Tibetans (CadetBlue ancestry) and Tai–Kadai-speaking populations (CadetBlue ancestry) can be separated from other East Asians. Tibetan–Burman-speaking Yi, Tujia and Bai have been separated from Tibetans and have similar ancestry profile with geographical neighbors and other Sinitic or Hmong–Mien-speaking populations. At $k=5$~8, no further substructure is observed in Tai–Kadai-speaking Zhuang, Dong and Bouyei, but a slight difference of ancestry component composition still exists among them. Our observed ancestry component distribution may reflect the recent genetic admixture or shared ancestry between Europeans and Asians (especially for Turkic populations), which we note may not be interpreted as informative evidence for human population migrations.

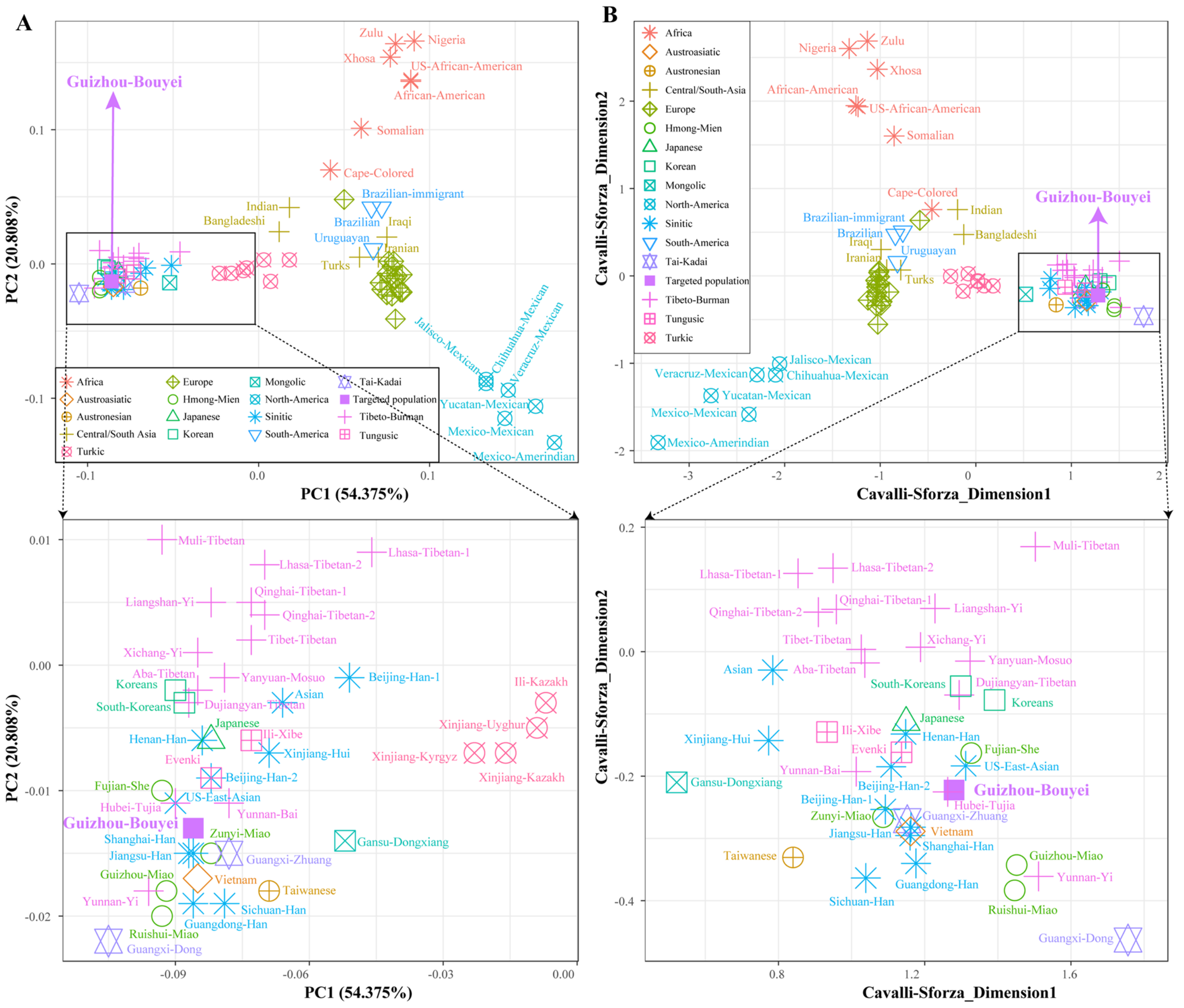

Fig. 4 Genetic similarities and differences between Tai–Kadai-speaking population and other 83 worldwide reference populations revealed by allele frequency distribution. **a** PCA results of the two-dimensional plots of the first two components; **b** patterns of population genetic relationship visualized by multidimensional scaling plots based on the Cavalli-Sforza genetic distances

Genetic affinity among 84 worldwide populations via allele frequency correlations

To comprehensively and delicately characterize the genetic landscape of Guizhou Bouyei and more geographically and linguistically structured reference populations, we assembled and conducted a frequency-based population comparison employing one new dataset composing of 14,303 subjects from 84 worldwide populations, including seven Africans, nine North Americans, 17 Europeans, five Central and South Asians, and other 46 Asian populations from Tai–Kadai, Austroasiatic, Austronesian, Hmong–Mien, Japonic, Koreanic, Sinitic, Tibeto-Burman, Mongolic, Tungusic, Turkic language families or groups. We initially carried out the PCA to characterize the genetic variability. The top five components capture a total of 89.421% variance (Fig. 4a, and Fig. S3a and b, PC1, 54.375%; PC2, 20.808%; PC3, 8.908%; PC4, 3.283% and PC5, 2.047%). A “Y” shape of the distribution patterns of population genetic relationships is identified (Fig. 4a). Three terminal points of “Y” patterns are, respectively, localized by Africans, Americans and Asians in clockwise order. PC1 captures the genetic differentiation between Americans, Asians, and Turkic-speaking populations with others, and PC2 can successfully separate Africans, Europeans and Americans. PC3 reflects the European and Asian cline of genetic differences and can also distinguish Turkic-speaking populations from other groups (Fig. S3a, b). Tai–Kadai-speaking Bouyei fall into an East

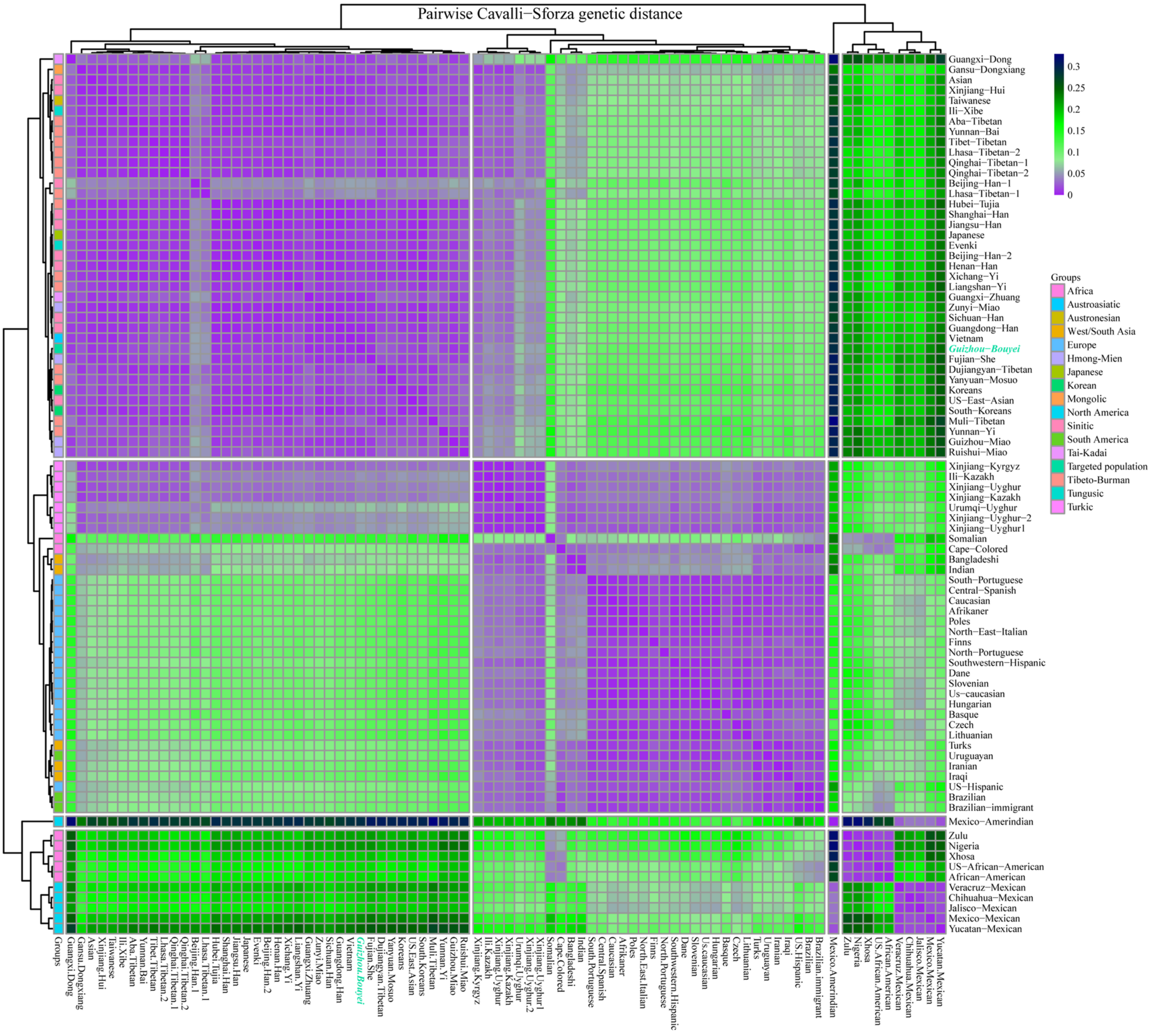

Fig. 5 Genetic homogeneity and heterogeneity among 84 populations displayed by the heatmap of the pairwise Cavalli-Sforza genetic distances

Asian cluster in the random combination of the first three components. Guizhou Miao is genetically close to Guangxi Zhuang, Zunyi Miao, Shanghai Han, Jiangsu Han, and Hubei Tujia. We subsequently explored patterns of genetic differentiation in Guizhou Bouyei and surrounding regions, as well as other continental references by estimating three popularly used pairwise genetic distances (Nei, Reynolds and Cavalli-Sforza, as shown in Fig. 5, Supplementary Figs. S4 and S5 and Tables S4–S6). The geographically close and linguistically similar populations show little genetic differentiation with Guizhou Bouyei. For example, Guangdong Han (0.0029 for Nei, 0.0042 for Cavalli-Sforza and 0.0043 for Reynolds), Zunyi Miao (0.0045, 0.0064 and 0.0064) and Hubei Tujia (0.0051, 0.0076 and 0.0074) show the smallest pairwise genetic distances with Guizhou Bouyei. Similar results are observed in Tai–Kadai-speaking Guangxi Zhuang, which is linguistically similar to Bouyei. Our pairwise Cavalli-Sforza genetic distance data, in the form of a heatmap plot, reveal four affinity clusters of low genetic differentiation, encompassing bio-geographically adjoining groups: the Europe/Turkic-speaking Asians, Africa, Asia, and America (Fig. 5). Consistent findings are also evidenced by the heatmap constructed by the pairwise Nei's distance in Supplementary Fig. S4 and Reynolds distance in Supplementary Fig. S5. In Asia, Turkic-speaking Uyghur, Kazakh and Kyrgyz are obviously more distant from their adjacent East Asian populations, whereas other Chinese populations from different language families exhibit relatively low interpopulation pairwise genetic distances, which are in close agreement with results of the PCA.

Additionally, we set out to further dissect and visualize the genetic similarities and dissimilarities among 84 reference populations via classical multidimensional scaling analyses on the basis of the aforementioned three distances. Figure 4d presents the two-dimensional plots on the basis of the Cavalli-Sforza genetic distance. Six North American populations keep a strong genetic affinity and are localized in the upper right position of the plots, while three South American populations are distinct with North American and have an affinity with the Central/Southern Asians which are located in the center position. Seven African groups are placed in the upper middle position, and East Asians are placed in the left part of plots. Europeans and Turkic-speaking populations are located in the intermediate positions between the African cluster and the East Asian cluster. We also observe the genetic difference between the Tibetan–Burman-speaking populations and other populations belonging to Sinitic, Tai–Kadai, Hmong–Mien language families. Since the numbers of Austronesian- and Austroasiatic-speaking populations are relatively small, there is no strong evidence for the elucidation of the association between genetic similarity and linguistic affinity. Our observations are also supported by the similar results from Nei's genetic distance (Supplementary Fig. S6) and Reynolds's genetic distance (Supplementary Fig. S7), as well as by the findings in aforementioned population genetic relationship exploration on the basis of raw genotype data.

Moreover, the genetic differentiation and phylogenetic relationship between Bouyei and other 83 reference populations are finally explored and reconstructed via three neighbor-joining trees (Fig. 6 and Supplementary Fig. S8). Three main genetic affinity clusters and several obvious subclusters within geographical or linguistic divisions can be clearly observed: North American cluster, European cluster, African cluster, Turkic-speaking population cluster, Tibeto-Burman-speaking population cluster and other admixture clusters. Bouyei samples cluster closely with Vietnamese and Zhuang populations. These results are consistent with the previous estimates, for instances, based on autosomal SNP and Y-chromosomal STRs (He et al. 2018b; Zou et al. 2018).

Discussion

Forensic related InDel markers can provide investigative clues when the samples collected from the crime scene are highly degraded or the cases with mutations. The forensic reference databases of the Investigator DIPplex amplification system have been subsequently genotyped and reported in Africans (Hefke et al. 2015), Europeans (Kis et al. 2012), South Asians (Akhteruzzaman et al. 2013), Central Asians (Poulsen et al. 2015), East Asians (Wang et al. 2014; He et al. 2019) and Americans (Martinez-Cortes et al. 2016). However, the forensic efficiency and reference database of this DIPplex panel in the Guizhou Bouyei remain uncharacterized. In the present study, we have investigated the genetic diversity, forensic allele frequency and corresponding statistical parameters of 30 InDels in 205 Guizhou Bouyei individuals. We have also explored the genetic relationships among the contemporary Tai–Kadai-speaking Bouyei in the context of worldwide populations based on both the raw genotype and frequency-based inference. The cumulative power of discrimination (CPD) and the combined power of exclusion (CPE) of those 30 markers are 0.99999999997 and 0.9841, respectively. Meng et al. once investigated the genetic polymorphisms of InDels in Chinese Xibe ethnic group and got similar forensic efficiency (CPE 0.9867 and CPD 0.9999999999902) (Meng et al. 2015). Other genetic analyses based on the obtained genotype data from other ethnolinguistically diverse populations, for instance, Yunnan Yi (Zhang et al. 2015), Tibetan (Guo et al. 2016), Hubei Tujia (Shen et al. 2016), Gansu Dongxiang (Zhu et al. 2018), Xinjiang Kazakh (Kong et al. 2017), Uyghur (Mei et al. 2016) and Hui (Xie et al. 2018), consistently obtained an expected discrimination and exclusion powers. Our findings that combined aforementioned investigated results indicated that the 30 InDel forensic identification system can be used in forensic applications. We also observed the power of forensic ancestry inference and population stratification using this DIP panel, like the forensic efficiency evaluation of the precision ancestry panel (He et al. 2018a). Recently, autosomal tri-allelic InDels and multi-allelic InDels have also been evidenced to have great considerable powers in forensic personal identification, parentage testing and even ancestry inference (Sun et al. 2016; Zhang et al. 2018; Zhao et al. 2018). Thus, more ancestry or identity informative InDel panel focusing on strong identification power in regional populations with continental discrimination should be developed and validated, especially for China with abundant linguistic, ethnic, cultural and genetic diversity.

Historical literature recorded that Bouyei and Zhuang were linked together and were referred to as Liliao, Manliao or Yiliao during the Wei dynasty to Tang Dynasty approximately 1400 years ago. The Bouyei was divided and called Zhongjia in the five Dynasties. Since then, the cultural and genetic differences between Bouyei and Zhuang population grew bigger due to the considerable changes in geographical and climatic conditions in the plains of Guizhou, the availability of resources and economic development, the ways of agriculture technology, and government local system reform. Besides, the ancestors of Bouyei and Zhuang may have admixed with their geographical neighbors in historic and prehistoric times. Previous genetic studies carried out

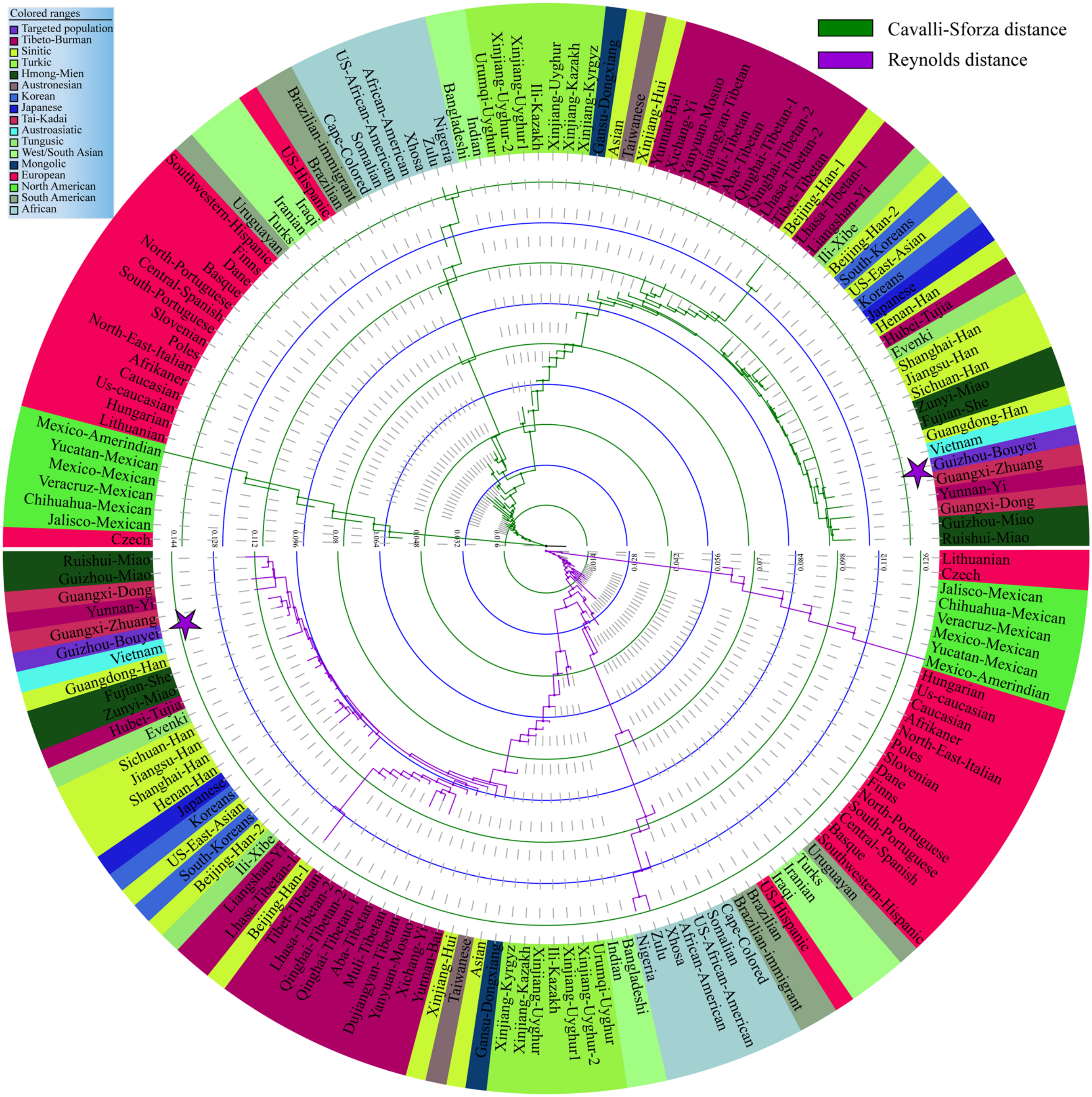

Fig. 6 Phylogenetic relationships among 84 worldwide populations reconstructed on the basis of the pairwise Cavalli-Sforza and Reynolds genetic distances

to date based on the human leukocyte antigen class I polymorphisms, biparental and Y-chromosomal highly mutated STRs examined the genetic diversity of Bouyei ethnic group and suggested that the Bouyei is genetically closer to the geographical adjacent Miao and Sui than other groups (Chen et al. 2007; Chen et al. 2018; Ren et al. 2018). In this study, we merged two large datasets, respectively, based on raw genotype data and allele frequency distribution of InDels from worldwide populations to characterize a fine-grained pattern of human genetic diversity of Bouyei and genetic relationships among our reference populations. Our examined patterns of human population substructure revealed significant genetic differences between continental populations and genetic homogeneity within the geographically or linguistically close populations. Guizhou Bouyei has a close genetic relationship with other Tai–Kadai-speaking populations (Zhuang and Dong) and geographic neighbors (Miao, Han and Tujia). The genetic ancestry profile of

Tibeto-Burman and Turkic speakers in East Asian strongly correlates with language family classifications. However, a mosaic correlation exists among Hmong–Mien and Tai–Kadai populations due to a large number of migrations and admixture events, or the common origin, as well as altering the genetic makeup via the language shift and linguistic assimilation (Huang et al. 2018). These observed co-evolution between language and genetics along some extent of differences are also observed in the Austronesian, Turkic and Uralic language families (Yunusbayev et al. 2015; Hudjashov et al. 2017; Tambets et al. 2018).

In summary, we obtained the first batch of forensic reference genotype database, allele frequency and forensic parameters of 30 autosomal InDel markers in Tai–Kadai-speaking Guizhou Bouyei population. Our results from the investigation of forensic characteristics have demonstrated that this autosomal InDels panel can be used as a powerful tool for forensic individual identification and parentage testing in the Bouyei population. Subsequently, we explored the genetic relationship of the different language groups or geographically continental groups with respect to a worldwide context, and explored the genetic diversity of Bouyei ethnic group as well as estimated the gene flow and population interaction across different linguistic phyla in China. Guizhou Bouyei people are genetically closest to the geographically adjacent Guizhou Miao and linguistically close Guangxi Zhuang and Dong rather than other reference populations. Genetic ancestry of Tibeto-Burman and Turkic speakers in east Asia strongly correlates with the classifications of language families. However, no strong association is identified between Hmong–Mien and Tai–Kadai populations, which may be influenced by the asymmetric semipermeable genetic introgression via the process of language shift and linguistic assimilation. Further genetic studies based on the whole-genome high-density variations of modern Bouyei or ancient human remains in the Guizhou plain before the Tang dynasty would be helpful for exploring the origin and reconstructing the detailed Bouyei population history.

Author contributions GH wrote the manuscripts. CCW revised the manuscript. HZ, ZR surveyed and collected the samples. HZ, ZR performed extraction of the genomic DNA. GH, JG, FZ, XZ, HZ, QW, JJ, MY, ZZ, JZ and YN performed the analysis of population diversity and genetic structure. CCW, JH initiated and supervised the project.

Funding The work was funded by the National Natural Science Foundation of China (81601650, 31801040), Guizhou Province Engineering Technology Research Center Project (Qian High-Tech of Development and Reform Commission No. [2016]1345), Nanqiang Outstanding Young Talents Program of Xiamen University (X2123302), and Fundamental Research Funds for the Central Universities (ZK1144).

Compliance with ethical standards

Conflict of interest The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Ethics statement We followed the recommendations of the “Medical Ethics Committee of Guizhou Medical University, China” in this study. All DNA samples are collected from unrelated donors with written informed consent. The protocol has been approved by the “Medical Ethics Committee of Guizhou Medical University” and in accordance with the Declaration of Helsinki.

References

Akhteruzzaman S, Das SA, Hosen I, Ferdous A (2013) Genetic polymorphism of 30 InDel markers for forensic use in Bangladeshi population. Forensic Sci Int Genet Suppl Ser 4:e348–e349

Barton L, Newsome SD, Chen FH, Wang H, Guilderson TP, Bettinger RL (2009) Agricultural origins and the isotopic identity of domestication in northern China. Proc Natl Acad Sci USA 106:5523–5528

Chen S, Ren X, Liu Y, Hu Q, Hong W, Xu A (2007) Human leukocyte antigen class I polymorphism in Miao, Bouyei, and Shui ethnic minorities of Guizhou, China. Hum Immunol 68:928–933

Chen P, He G, Zou X, Zhang X, Li J, Wang Z, Gao H, Luo L, Zhang Z, Yu J, Han Y (2018) Genetic diversities and phylogenetic analyses of three Chinese main ethnic groups in southwest China: a Y-Chromosomal STR study. Sci Rep 8:15339

Earl DA, vonHoldt BM (2011) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour 4:359–361

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

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

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

Fondevila M, Phillips C, Santos C, Pereira R, Gusmao L, Carracedo A, Butler JM, Lareu MV, Vallone PM (2012) Forensic performance of two insertion-deletion marker assays. Int J Legal Med 126:725–737

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

Guo Y, Shen C, Meng H, Dong Q, Kong T, Yang C, Wang H, Jin R, Zhu B (2016) Population differentiations and phylogenetic analysis of Tibet and Qinghai Tibetan groups based on 30 InDel Loci. DNA Cell Biol 35:787–794

Hansen J (2005) Using SPSS for windows and macintosh: analyzing and understanding data. Am Stat 59:113

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

He G, Wang Z, Wang M, Zou X, Liu J, Wang S, Hou Y (2018b) Genetic variations and forensic characteristics of Han Chinese population residing in the Pearl River Delta revealed by 23 autosomal STRs. Mol Biol Rep 45:1125–1133

He G, Wang Z, Zou X, Wang M, Liu J, Wang S, Ye Z, Chen P, Hou Y (2019) Tai-Kadai-speaking Gelao population: forensic features, genetic diversity and population structure. Forensic Sci Int Genet 40:e231–e239

Hefke G, Davison S, D'Amato ME (2015) Forensic performance of investigator DIPplex InDels genotyping kit in native, immigrant, and admixed populations in South Africa. Electrophoresis 36:3018–3025

Huang X, Zhou Q, Bin X, Lai S, Lin C, Hu R, Xiao J, Luo D, Li Y, Wei LH, Yeh HY, Chen G, Wang CC (2018) The genetic assimilation in language borrowing inferred from Jing People. Am J Phys Anthropol 166:638–648

Hudjashov G, Karafet TM, Lawson DJ, Downey S, Savina O, Sudoyo H, Lansing JS, Hammer MF, Cox MP (2017) Complex Patterns of Admixture across the Indonesian Archipelago. Mol Biol Evol 34:2439–2452

Inacio A, Costa HA, da Silva CV, Ribeiro T, Porto MJ, Santos JC, Igrejas G, Amorim A (2017) Study of InDel genetic markers with forensic and ancestry informative interest in PALOP's immigrant populations in Lisboa. Int J Legal Med 131:657–660

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23:1801–1806

Kalinowski ST (2002) Evolutionary and statistical properties of three genetic distances. Mol Ecol 11:1263–1273

Kis Z, Zalan A, Volgyi A, Kozma Z, Domjan L, Pamjav H (2012) Genome deletion and insertion polymorphisms (DIPs) in the Hungarian population. Forensic Sci Int Genet 6:e125–e126

Kong T, Chen Y, Guo Y, Wei Y, Jin X, Xie T, Mu Y, Dong Q, Wen S, Zhou B, Zhang L, Shen C, Zhu B (2017) Autosomal InDel polymorphisms for population genetic structure and differentiation analysis of Chinese Kazak ethnic group. Oncotarget 8:56651–56658

Kovach WL (2007) MVSP-A multivariate statistical package for windows, ver. 3.1. Kovach Computing Services, Pentraeth, Wales

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

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

Lipson M, Skoglund P, Spriggs M, Valentin F, Bedford S, Shing R, Buckley H, Phillip I, Ward GK, Mallick S, Rohland N, Broomandkhoshbacht N, Cheronet O, Ferry M, Harper TK, Michel M, Oppenheimer J, Sirak K, Stewardson K, Auckland K, Hill AVS, Maitland K, Oppenheimer SJ, Parks T, Robson K, Williams TN, Kennett DJ, Mentzer AJ, Pinhasi R, Reich D (2018b) Population turnover in remote Oceania shortly after initial settlement. Curr Biol 28(1157–1165):e1157

Lu D, Lou H, Yuan K, Wang X, Wang Y, Zhang C, Lu Y, Yang X, Deng L, Zhou Y, Feng Q, Hu Y, Ding Q, Yang Y, Li S, Jin L, Guan Y, Su B, Kang L, Xu S (2016) Ancestral origins and genetic history of Tibetan highlanders. Am J Hum Genet 99:580–594

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

Martinez-Cortes G, Garcia-Aceves M, Favela-Mendoza AF, Munoz-Valle JF, Velarde-Felix JS, Rangel-Villalobos H (2016) Forensic parameters of the Investigator DIPplex kit (Qiagen) in six Mexican populations. Int J Legal Med 130:683–685

McColl H, Racimo F, Vinner L, Demeter F, Gakuhari T, Moreno-Mayar JV, van Driem G, Gram Wilken U, Seguin-Orlando A, de la Fuente Castro C, Wasef S, Shoocongdej R, Souksavatdy V, Sayavongkhamdy T, Saidin MM, Allentoft ME, Sato T, Malaspinas AS, Aghakhanian FA, Korneliussen T, Prohaska A, Margaryan A, de Barros Damgaard P, Kaewsutthi S, Lertrit P, Nguyen TMH, Hung HC, Minh Tran T, Nghia Truong H, Nguyen GH, Shahidan S, Wiradnyana K, Matsumae H, Shigehara N, Yoneda M, Ishida H, Masuyama T, Yamada Y, Tajima A, Shibata H, Toyoda A, Hanihara T, Nakagome S, Deviese T, Bacon AM, Duringer P, Ponche JL, Shackelford L, Patole-Edoumba E, Nguyen AT, Bellina-Pryce B, Galipaud JC, Kinaston R, Buckley H, Pottier C, Rasmussen S, Higham T, Foley RA, Lahr MM, Orlando L, Sikora M, Phipps ME, Oota H, Higham C, Lambert DM, Willerslev E (2018) The prehistoric peopling of Southeast Asia. Science 361:88–92

Mei T, Shen CM, Liu YS, Meng HT, Zhang YD, Guo YX, Dong Q, Wang XX, Yan JW, Zhu BF, Zhang LP (2016) Population genetic structure analysis and forensic evaluation of Xinjiang Uigur ethnic group on genomic deletion and insertion polymorphisms. Springerplus 5:1087

Meng HT, Zhang YD, Shen CM, Yuan GL, Yang CH, Jin R, Yan JW, Wang HD, Liu WJ, Jing H, Zhu BF (2015) Genetic polymorphism analyses of 30 InDels in Chinese Xibe ethnic group and its population genetic differentiations with other groups. Sci Rep 5:8260

Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, Devine SE (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res 16:1182–1190

Nei M (1978) The theory of genetic distance and evolution of human races. Jinrui Idengaku Zasshi 23:341–369

Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E (2017) Tracing the peopling of the world through genomics. Nature 541:302–310

Olalde I, Brace S, Allentoft ME, Armit I, Kristiansen K, Booth T, Rohland N, Mallick S, Szecsenyi-Nagy A, Mittnik A, Altena E, Lipson M, Lazaridis I, Harper TK, Patterson N, Broomandkhoshbacht N, Diekmann Y, Faltyskova Z, Fernandes D, Ferry M, Harney E, de Knijff P, Michel M, Oppenheimer J, Stewardson K, Barclay A, Alt KW, Liesau C, Rios P, Blasco C, Miguel JV, Garcia RM, Fernandez AA, Banffy E, Bernabo-Brea M, Billoin D, Bonsall C, Bonsall L, Allen T, Buster L, Carver S, Navarro LC, Craig OE, Cook GT, Cunliffe B, Denaire A, Dinwiddy KE, Dodwell N, Ernee M, Evans C, Kucharik M, Farre JF, Fowler C, Gazenbeek M, Pena RG, Haber-Uriarte M, Haduch E, Hey G, Jowett N, Knowles T, Massy K, Pfrengle S, Lefranc P, Lemercier O, Lefebvre A, Martinez CH, Olmo VG, Ramirez AB, Maurandi JL, Majo T, McKinley JI, McSweeney K, Mende BG, Modi A, Kulcsar G, Kiss V, Czene A, Patay R, Endrodi A, Kohler K, Hajdu

T, Szeniczey T, Dani J, Bernert Z, Hoole M, Cheronet O, Keating D, Veleminsky P, Dobes M, Candilio F, Brown F, Fernandez RF, Herrero-Corral AM, Tusa S, Carnieri E, Lentini L, Valenti A, Zanini A, Waddington C, Delibes G, Guerra-Doce E, Neil B, Brittain M, Luke M, Mortimer R, Desideri J, Besse M, Brucken G, Furmanek M, Haluszko A, Mackiewicz M, Rapinski A, Leach S, Soriano I, Lillios KT, Cardoso JL, Pearson MP, Wlodarczak P, Price TD, Prieto P, Rey PJ, Risch R, Rojo Guerra MA, Schmitt A, Serralongue J, Silva AM, Smrcka V, Vergnaud L, Zilhao J, Caramelli D, Higham T, Thomas MG, Kennett DJ, Fokkens H, Heyd V, Sheridan A, Sjogren KG, Stockhammer PW, Krause J, Pinhasi R, Haak W, Barnes I, Lalueza-Fox C, Reich D (2018) The Beaker phenomenon and the genomic transformation of northwest Europe. Nature 555:190–196

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

Phillips C (2015) Forensic genetic analysis of bio-geographical ancestry. Forensic Sci Int Genet 18:49–65

Poulsen L, Farzad MS, Borsting C, Tomas C, Pereira V, Morling N (2015) Population and forensic data for three sets of forensic genetic markers in four ethnic groups from Iran: Persians, Lurs, Kurds and Azeris. Forensic Sci Int Genet 17:43–46

Ren Z, Zhang H, Liu Y, Wang Q, Wang J, Huang J (2018) Population genetic data of 22 autosomal STRs in Guizhou Bouyei population, Southwestern China. Forensic Sci Int Genet 33:e11–e12

Reynolds J, Weir BS, Cockerham CC (1983) Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics 105:767–779

Rosenberg NA (2004) DISTRUCT: a program for the graphical display of population structure. Mol Ecol Notes 4:137–138

Shen C, Zhu B, Yao T, Li Z, Zhang Y, Yan J, Wang B, Bie X, Tai F (2016) A 30-InDel assay for genetic variation and population structure analysis of Chinese Tujia group. Sci Rep 6:36842

Siska V, Jones ER, Jeon S, Bhak Y, Kim HM, Cho YS, Kim H, Lee K, Veselovskaya E, Balueva T, Gallego-Llorente M, Hofreiter M, Bradley DG, Eriksson A, Pinhasi R, Bhak J, Manica A (2017) Genome-wide data from two early Neolithic East Asian individuals dating to 7700 years ago. Sci Adv 3:e1601877

Su B, Xiao JH, Underhill P, Deka R, Zhang WL, Akey J, Huang W, Shen D, Lu D, Luo JC, Chu JY, Tan JZ, Shen PD, Davis R, Cavalli-Sforza L, Chakraborty R, Xiong MM, Du RF, Oefner P, Chen Z, Jin L (1999) Y-chromosome evidence for a northward migration of modern humans into eastern Asia during the last Ice Age. Am J Hum Genet 65:1718–1724

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH, Konkel MK, Malhotra A, Stutz AM, Shi X, Casale FP, Chen J, Hormozdiari F, Dayama G, Chen K, Malig M, Chaisson MJP, Walter K, Meiers S, Kashin S, Garrison E, Auton A, Lam HYK, Mu XJ, Alkan C, Antaki D, Bae T, Cerveira E, Chines P, Chong Z, Clarke L, Dal E, Ding L, Emery S, Fan X, Gujral M, Kahveci F, Kidd JM, Kong Y, Lameijer EW, McCarthy S, Flicek P, Gibbs RA, Marth G, Mason CE, Menelaou A, Muzny DM, Nelson BJ, Noor A, Parrish NF, Pendleton M, Quitadamo A, Raeder B, Schadt EE, Romanovitch M, Schlattl A, Sebra R, Shabalin AA, Untergasser A, Walker JA, Wang M, Yu F, Zhang C, Zhang J, Zheng-Bradley X, Zhou W, Zichner T, Sebat J, Batzer MA, McCarroll SA, Genomes Project C, Mills RE, Gerstein MB, Bashir A, Stegle O, Devine SE, Lee C, Eichler EE, Korbel JO (2015) An integrated map of structural variation in 2504 human genomes. Nature 526:75–81

Sun K, Ye Y, Luo T, Hou Y (2016) Multi-InDel analysis for ancestry inference of sub-populations in China. Sci Rep 6:39797

Tambets K, Yunusbayev B, Hudjashov G, Ilumae AM, Rootsi S, Honkola T, Vesakoski O, Atkinson Q, Skoglund P, Kushniarevich A, Litvinov S, Reidla M, Metspalu E, Saag L, Rantanen T, Karmin M, Parik J, Zhadanov SI, Gubina M, Damba LD, Bermisheva M, Reisberg T, Dibirova K, Evseeva I, Nelis M, Klovins J, Metspalu A, Esko T, Balanovsky O, Balanovska E, Khusnutdinova EK, Osipova LP, Voevoda M, Villems R, Kivisild T, Metspalu M (2018) Genes reveal traces of common recent demographic history for most of the Uralic-speaking populations. Genome Biol 19:139

Wang Z, Zhang S, Zhao S, Hu Z, Sun K, Li C (2014) Population genetics of 30 insertion-deletion polymorphisms in two Chinese populations using Qiagen Investigator(R) DIPplex kit. Forensic Sci Int Genet 11:e12–e14

Weber JL, David D, Heil J, Fan Y, Zhao C, Marth G (2002) Human diallelic insertion/deletion polymorphisms. Am J Hum Genet 71:854–862

Xie T, Guo Y, Chen L, Fang Y, Tai Y, Zhou Y, Qiu P, Zhu B (2018) A set of autosomal multiple InDel markers for forensic application and population genetic analysis in the Chinese Xinjiang Hui group. Forensic Sci Int Genet 35:1–8

Xu S, Jin L (2008) A genome-wide analysis of admixture in Uyghurs and a high-density admixture map for disease-gene discovery. Am J Hum Genet 83:322–336

Xu S, Huang W, Qian J, Jin L (2008) Analysis of genomic admixture in Uyghur and its implication in mapping strategy. Am J Hum Genet 82:883–894

Yang X, Wan Z, Perry L, Lu H, Wang Q, Zhao C, Li J, Xie F, Yu J, Cui T, Wang T, Li M, Ge Q (2012) Early millet use in northern China. Proc Natl Acad Sci USA 109:3726–3730

Yang MA, Gao X, Theunert C, Tong H, Aximu-Petri A, Nickel B, Slatkin M, Meyer M, Paabo S, Kelso J, Fu Q (2017) 40,000-year-Old Individual from Asia Provides insight into early population structure in Eurasia. Curr Biol 27(3202–3208):e3209

Yunusbayev B, Metspalu M, Metspalu E, Valeev A, Litvinov S, Valiev R, Akhmetova V, Balanovska E, Balanovsky O, Turdikulova S, Dalimova D, Nymadawa P, Bahmanimehr A, Sahakyan H, Tambets K, Fedorova S, Barashkov N, Khidiyatova I, Mihailov E, Khusainova R, Damba L, Derenko M, Malyarchuk B, Osipova L, Voevoda M, Yepiskoposyan L, Kivisild T, Khusnutdinova E, Villems R (2015) The genetic legacy of the expansion of Turkic-speaking nomads across Eurasia. PLoS Genet 11:e1005068

Zhang YD, Shen CM, Jin R, Li YN, Wang B, Ma LX, Meng HT, Yan JW, Dan Wang H, Yang ZL, Zhu BF (2015) Forensic evaluation and population genetic study of 30 insertion/deletion polymorphisms in a Chinese Yi group. Electrophoresis 36:1196–1201

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

Zhao X, Chen X, Zhao Y, Zhang S, Gao Z, Yang Y, Wang Y, Zhang J (2018) Construction and forensic genetic characterization of 11 autosomal haplotypes consisting of 22 tri-allelic InDels. Forensic Sci Int Genet 34:71–80

Zhu B, Lan Q, Guo Y, Xie T, Fang Y, Jin X, Cui W, Chen C, Zhou Y, Li X (2018) Population genetic diversity and clustering analysis for Chinese Dongxiang group with 30 autosomal InDel loci simultaneously analyzed. Front Genet 9:279

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

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.