Frontiers in Bioscience Landmark

Front. Biosci. (Landmark Ed) **2023**; 28(6): 114
https://doi.org/10.31083/j.fbl2806114

Original Research

Forensic Characterization and Genetic Portrait of the Gannan Tibetan Ethnic Group via 165 AI-SNP Loci

Wei Cui[1,2], Man Chen[1,2], Hongbing Yao[3], Qing Yang[4], Liu Liu[1], Xiaole Bai[1], Ling Chen[1,*], Bofeng Zhu[1,2,5,*]

[1]Guangzhou Key Laboratory of Forensic Multi-Omics for Precision Identification, School of Forensic Medicine, Southern Medical University, 510515 Guangzhou, Guangdong, China
[2]Microbiome Medicine Center, Department of Laboratory Medicine, Zhujiang Hospital, Southern Medical University, 510280 Guangzhou, Guangdong, China
[3]Belt and Road Research Center for Forensic Molecular Anthropology, Key Laboratory of Evidence Science of Gansu Province, Gansu University of Political Science and Law, 730070 Lanzhou, Gansu, China
[4]Genetic Sciences Group, Thermo Fisher Scientific (China) Inc. Guangzhou Branch, 510005 Guangzhou, Guangdong, China
[5]Key Laboratory of Shaanxi Province for Craniofacial Precision Medicine Research, College of Stomatology, Xi'an Jiaotong University, 710004 Xi'an, Shaanxi, China
*Correspondence: lingpzy@163.com (Ling Chen); zhubofeng7372@126.com (Bofeng Zhu)
Academic Editor: Said El Shamieh
Submitted: 25 October 2022 Revised: 12 January 2023 Accepted: 31 January 2023 Published: 14 June 2023

Abstract

Background: The Tibetan group is one of the oldest Sino-Tibetan ethnic groups. The origin, migration as well as the genetic background of Tibetans have become the research hotspots in the field of forensic genetics. The use of ancestry informative markers (AIMs) allows the investigation of the genetic background of the Gannan Tibetan group. **Methods**: In this study, the 165 ancestry informative single nucleotide polymorphism (AI-SNP) loci included in the Precision ID Ancestry Panel were used to genotype 101 Gannan Tibetans using the Ion S5 XL system. The forensic statistical parameters of 165 AI-SNP in the Gannan Tibetan group were calculated. Population genetic analyses including *Nei's* genetic distances, phylogenetic analyses, pairwise fixation index, principal component analyses and population ancestry composition analyses were also conducted to evaluate the genetic relationships between the Gannan Tibetan group and other reference populations. **Results**: Forensic parameters of the 165 AI-SNP loci indicated that not all of the SNPs showed high genetic polymorphisms in the Gannan Tibetan group. Population genetic analyses indicated that the Gannan Tibetan group had close genetic affinities with East Asian populations, especially with the groups residing in its neighboring geographical regions. **Conclusions**: The 165 AI-SNP loci in the Precision ID Ancestry Panel showed high ancestral prediction powers for different continental populations. When trying to predict the ancestral information of East Asian subpopulations using this panel, the prediction results are not particularly accurate. The 165 AI-SNP loci showed varying degrees of genetic polymorphisms in the Gannan Tibetan group, and the combined use of these loci could be an effective tool in the forensic individual identification and parentage testing of this group. The Gannan Tibetan group has close genetic affinities with East Asian populations compared with other reference populations, especially tighter genetic relationships with the groups residing in its neighboring geographical regions.

Keywords: ancestry informative SNP; forensic genetic analyses; massively parallel sequencing; Gannan Tibetan group; precision ID ancestry panel

1. Introduction

The Tibetan group, indigenous group of the Qinghai-Tibet Plateau, is one of the 56 ethnic groups with a long history in China. Chinese Tibetan group mainly settled in the Tibet Autonomous Region, Qinghai, Western Sichuan, Yunnan, and the Gansu provinces [1,2]. According to the seventh national population census in 2020, the population of the Tibetans is more than seven millions [3]. The Tibetan language is a branch of the Tibeto-Burman family of the Sino-Tibetan language. Chinese Tibetan group could be divided into three main subgroups (U-Tsang, Khams and Amdo) in terms of their geographical and cultural differences [4]. The Gansu Tibetans, belonging to a branch of the Amdo Tibetans, mainly reside in Gannan Tibetan Autonomous Prefecture, which is located on the northeastern edge of the Qinghai-Tibet Plateau, and at the border of the Aba Tibetan Qiang Autonomous Prefecture of the Sichuan province [5]. In the past decade, many researchers devoted their efforts to uncovering the origin, migration, genetic admixture as well as high-altitude adaptation of the Tibetan group [1,6–8]. However, these issues have been long-standing heated arguments because of limited genetic data of ancient and modern Tibetans.

In recent years, ancestry inference of unknown DNA donor found at the crime scene has become a new tool to help solve crime [9]. Ancestry informative marker (AIM) refers to a kind of genetic markers with significant frequency variations among different populations. With the

Copyright: © 2023 The Author(s). Published by IMR Press.
This is an open access article under the CC BY 4.0 license.

Publisher's Note: IMR Press stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

technological developments of next generation sequencing (NGS) and the reduction of the sequencing cost, several biogeographical ancestry inference panels based on single nucleotide polymorphism (SNP), deletion-insertion polymorphism (DIP) and multi-allelic haplotype markers have been constructed [10–12]. Nowadays, ancestry information inference has wide applications not only in crime investigations, but also in the identifications of skeletal remains of missing persons, population substructure studies, population genetic investigations, and the disease susceptibility studies in different populations [13–16].

With the exception of AI-SNP panels recently developed for East Asian subpopulation structure [17–21], few commercial kits are available in the forensic genetic field. ForenSeq™ DNA Signature Prep Kit and Precision ID Ancestry Panel are both commonly used in forensic DNA laboratory. According to related validation study, the 165 AI-SNP loci of the Precision ID Ancestry Panel can be used to discriminate the ancestal origins of major populations including Africa, Europe, Southwest Asia, South Asia, East Asia, Oceania and Americas [22]. Population genetic evaluations based on Precision ID Ancestry Panel indicated that this panel could also be used to analyze the genetic background of some subpopulations [23,24]. In this study, we systematically evaluated the performance of the Precision ID Ancestry Panel in 101 Tibetan individuals from Gannan Tibetan Autonomous Prefecture of Chinese Gansu province. The present results provided the raw data on genetic variations at 165 SNP loci in Gannan Tibetan ethnic minority, the relevant forensic parameters for individual discrimination in this region and indicated the genetic relationships between Gannan Tibetan group and reference groups.

2. Materials and Methods

2.1 Ethical Statement and Sample Collection

The present study has been approved by the ethics committee of Xi'an Jiaotong University Health Science Center and Southern Medical University (Ethical Approval Number: 2019-1039). This research was conducted in accordance with the ethical principle for medical research involving human subjects recommended by the World Medical Association Declaration of Helsinki. Sample collection and the following sequencing experiments were performed in strict compliance with the ethical regulations. Peripheral venous blood samples were collected from 101 unrelated volunteers who lived in the Gannan Tibetan Autonomous Prefecture in the Gansu province, China for at least three generations. The blood samples were dried on the FTA cards, and then stored at room temperature. Each volunteer signed an informed consent form before the sample collection.

2.2 Genomic DNA Extraction and DNA Quantification

For each sample, five pieces of 1.0 mm^2 bloodstain were used to extract the genomic DNA using the QIAamp® DNA Investigator kit (QIAGEN, Hilden, Germany) according to the manufacturer's instructions [25]. The concentration of genomic DNA was then measured by quantitative fluorescence using the Qubit™ dsDNA HS Assay kit (Thermo Fisher Scientific Inc, Waltham, MA, USA) and the Qubit 4.0 Fluorometer (Thermo Fisher Scientific Inc, Waltham, MA, USA) following the manufacturer's recommendations [26].

2.3 Library Preparation and Quantification

The Precision ID DL8 Kit and the Precision ID Ancestry Panel (Thermo Fisher Scientific Inc, Waltham, MA, USA) were used to prepare DNA library on the automated Ion Chef™ system (Thermo Fisher Scientific Inc, Waltham, MA, USA). For each sample, 15 µL of DNA (1 ng in total) was pipetted into A1 to A8 wells of a Precision ID DL8 IonCode™ Barcode Adapters plate. And twenty-one cycles were used to amplify the target regions. The concentrations of sample libraries were measured by the Ion Library TaqMan® Quantitation Kit (Thermo Fisher Scientific Inc, Waltham, MA, USA) on the QuantStudio™ 5 real-time PCR system (Thermo Fisher Scientific Inc, Waltham, MA, USA) following the manufacturer's instructions. Before template preparation, all DNA libraries were pooled in equimolar concentrations (33 pM).

2.4 Template Preparation and DNA Sequencing

Template preparation, enrichment of beads with template, and the sequence of template on beads were performed with the Ion Chef™ and Ion S5™ XL (Thermo Fisher Scientific Inc, Waltham, MA, USA) instruments according to the manufacturer's instructions. After template preparation, high-throughput sequencing was preformed using four pieces of Ion™ 530 chips (Thermo Fisher Scientific Inc, Waltham, MA, USA). Thirty-two samples were loaded on each Ion™ 530 chip. Positive DNA 007 and deionized water were used as positive and negative controls in each run.

2.5 Analyses of NGS Data

The analysis of primary sequence data was performed with the Torrent Suite Software (Thermo Fisher Scientific Inc, Waltham, MA, USA). All reads were aligned to hg19 reference sequences. Further sequence analyses were carried out using the HID-SNP Genotyper v4.2 plug-in and Converge™ Software (Thermo Fisher Scientific Inc, Waltham, MA, USA) with default analysis settings. Statistical analyses of read depths and balance ratios for the 165 AI-SNPs were calculated using the OriginPro 2021b software (version 9.8, OriginLab Corporation, Northampton, MA, USA).

2.6 Sanger Sequencing of Locus rs7722456

DNA samples with an 'A' allele at locus rs7722456 were further confirmed by Sanger sequencing. Forward primer (5′-GCTGACTCTAGCCCTTTGGG-3′) and reverse primer (5′-GGTGGGTCTTGTTTGGCATT-3′) were synthesized and then used to amplify the locus rs7722456. The PCR products were separated by the method of agarose gel electrophoresis, and then purified using the DiaSpin DNA Gel Extraction kit (Sangon Biotech, Guangzhou, China) following manufacturer's instructions. Sanger sequencing was conducted at rs7722456 locus by Sangon Biotech (Sangon Biotech, Guangzhou, China).

2.7 Reference Populations

In the present study, 164 AI-SNP data sets of 26 populations were acquired from the 1000 Genomes Project Phase III database (locus rs10954737 was excluded from the population genetic studies as its genotype data were not available in the 1000 Genomes Project) [27]. The 165 AIM-SNP data sets for 22 populations were obtained from HGDP-CEPH database [28]. SNP genotype data for nine populations (Sichuan Tibetan, Qinghai Tibetan and Liangshan Yi [29], Hainan Han, Hainan Li and Gelao [23], Wuzhong Hui [30], Chinese Kazak [31], and Basque [32]) were obtained from previously published literature. Detailed information, such as population names and their abbreviations, population sizes of all reference populations, is listed in **Supplementary Table 1**.

2.8 Population Genetic Analyses

The linkage disequilibrium (LD) test, allelic frequencies and forensic statistical parameters of the 165 AI-SNP loci in the Gannan Tibetan group were calculated using STRAF [33]. Tests of Hardy-Weinberg equilibrium (HWE) for 165 AI-SNP loci in Gannan Tibetan group were performed by the Arlequin software (version 3.5.1.2, Laurent Excoffier & Heidi Lischer, Berne, Switzerland) [34] . The heatmap for the minimal allelic frequencies (MAF) of 164 AI- SNP loci was created using the *R* software (version 4.0.5, R Foundation for Statistical Computing, Vienna, Austria). Pairwise fixation index (F_{ST}) values, which were used to measure the genetic distances of pairwise populations, were calculated between Gannan Tibetan group and reference populations using the Arlequin software (version 3.5.1.2, Excoffier & Lischer, Berne, Switzerland) [34]. Moreover, *Nei's* genetic distances (D_A) between the Gannan Tibetan group and other reference populations were calculated using the DISPAN program (Nei & Tajima & Tateno, Kanagawa, Japan) [35]. A neighbor-joining (NJ) phylogenetic tree was constructed on the basis of pairwise D_A values by the MEGA software (version 7, Sudhir Kumar & Glen Stecher & Koichiro Tamura, PA, USA) [36], and then visualized with the ggtree package (version 3.4.2, Guangchuang Yu, Guangdong, China) [37] from the *R* software (version 4.0.5, R Foundation for Statistical Computing, Vienna, Austria). Phylogenetic trees constructed using the maximum likelihood (ML) method were generated using the TreeMix software (version 1.13, Joseph K. Pickrell & Jonathan K. Pritchard, IL, USA), which assumed that 0–10 migration events happened among all the populations [38]. Principal component analyses (PCA) were also carried out to evaluate the genetic relationships between the Gannan Tibetan group and reference groups. PCAs at individual scale were performed based on the genotype data of 164 AI-SNP loci of Gannan Tibetan group and reference populations using the *R* software and the SmartPCA package implemented in the EIGENSOFT software (version 6.1.4, Alkes L Price & Nick J Patterson, Boston, MA, USA), respectively [39]. Population genetics and ancestry component analyses of the Gannan Tibetan group were performed using the model-based ADMIXTURE (version 1.3, David H Alexander & John Novembre & Kenneth Lange, Los Angeles, CA, USA) and STRUCTURE softwares (version 2.3.4, Pritchard & Stephens & Donnelly, Oxford, UK) [40,41] with the setting K values from 2 to 6. The CLUMPAK tool (http://clumpak.tau.ac.il/) was used to analyze the results of ADMIXTURE (version 1.3, David H Alexander & John Novembre & Kenneth Lange, Los Angeles, CA, USA).

3. Results

3.1 Assessment for the NGS Results

In this study, all 165 AI-SNP genotypes in 101 Gannan Tibetan individuals and positive control DNA were successfully generated using the Ion S5™ XL system. Four Ion™ 530 chips generated a total of 1.5~1.87 gigabases (Gb). Ion sphere particle (ISP) loading rate of each Ion™ 530 chip was more than 78%. As recommended by the manufacturer's instructions, total reads should be greater than 1 Gb; the ISP loading rate should be greater than 50%; and the total usable reads should be greater than 30%. These sequencing metrics all met the recommended values of the manufacturer's instructions. Read depths of 165 AI-SNP loci in 101 Gannan Tibetan individuals are shown in **Supplementary Fig. 1**. The minimum depth of coverage was 24×, which was noted at rs9845457 locus in sample GNZ13, while the maximum read depth was 8287×, which was seen at rs7657799 locus in GNZ54. The mean read depths for all 101 individuals ranged from 231 ± 87× (mean ± standard deviation) to 2932 ± 1359×. The balance ratios of all amplicons from the forward direction in the 101 samples are shown in Fig. 1, and the box plots revealed that the balance ratios of almost all amplicons were around 0.5.

Sixty-three individuals (accounting for 62.4% of total tested Gannan Tibetan individuals) were genotyped 'AT' at rs7722456. Genotypes of rs7722456 at these samples were further confirmed by Sanger sequencing. Results of Sanger sequencing showed that the genotypes in these 63 individuals were 'TT' at rs7722456 (**Supplementary Fig. 2**).

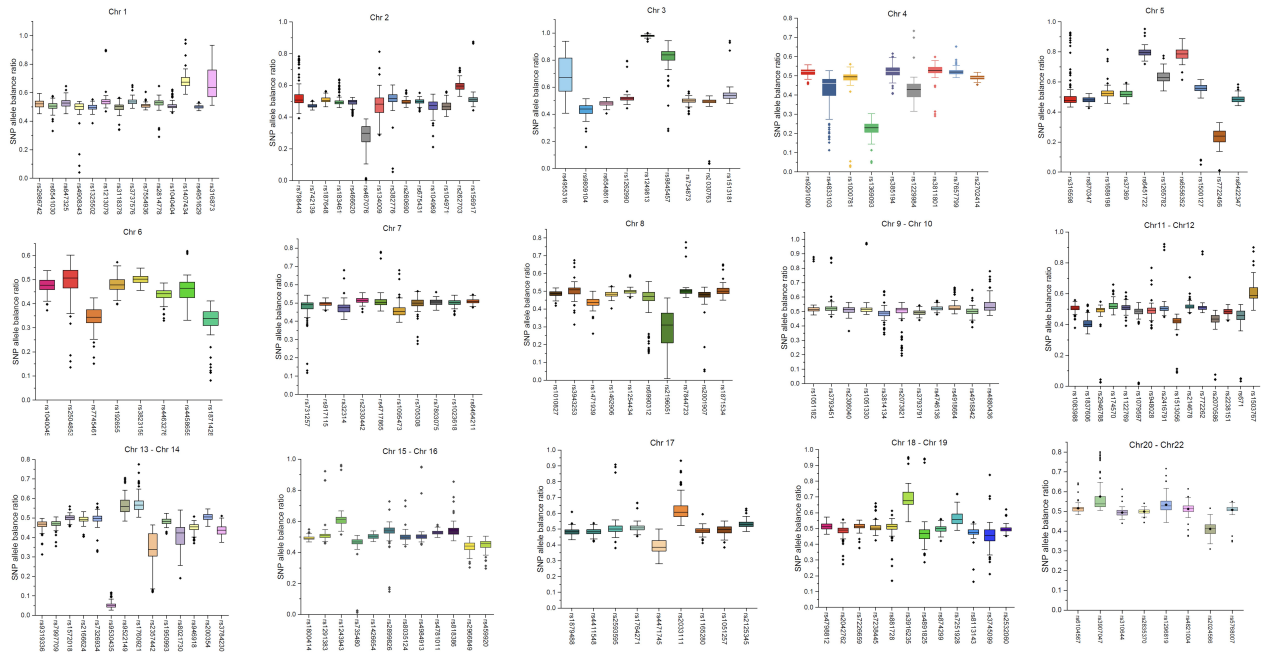

Fig. 1. Box plots of the balance ratios from forward direction of all SNP loci in 101 Gannan Tibetan individuals.

3.2 HWE and LD Tests of 165 AI-SNPs in the Gannan Tibetan Group

Before the population genetic investigations, the HWE tests of the 165 AI-SNPs were carried out in the Gannan Tibetan group, and the results are shown in **Supplementary Table 2**. The rs1462906, rs1871534, rs2814778, rs3916235, rs4880436, rs4891825, rs6754311, rs7326934 and rs7722456 of 165 AI-SNPs were excluded from the HWE tests because of their homozygous genotypes. The results of HWE tests showed that no significant deviations from HWE for the remaining 156 AI-SNPs were observed in the Gannan Tibetan group after sequential Bonferroni corrections ($p = 0.05/156 = 0.0003$).

The LD tests for pairwise SNP loci were performed based on the genotype data, and the *p*-values are shown in **Supplementary Table 3**. Similarly, rs1462906, rs1871534, rs2814778, rs3916235, rs4880436, rs4891825, rs6754311, rs7326934 and rs7722456 were also excluded from the LD tests because of their homozygous genotypes. After sequential Bonferroni corrections ($p = 0.05/24150 = 2.07 \times 10^{-6}$), all pairwise SNP loci reached to the state of linkage equilibrium in the studied Gannan Tibetan group except for the pair of rs1229984 and rs3811801 ($p = 8.72 \times 10^{-19}$).

3.3 Forensic Statistical Parameters and Allelic Frequency Distributions of 165 AI-SNPs in the Gannan Tibetan Group

Among the 165 AI-SNPs, nine loci (rs1462906, rs1871534, rs2814778, rs3916235, rs4880436, rs4891825, rs6754311, rs7326934 and rs7722456) were homozygous. Forensic statistical parameters including expected heterozygosity (H_{exp}), matching probability (MP), power of discrimination (PD), polymorphism information content (PIC), observed heterozygosity (H_{obs}), probability of exclusion (PE) and typical paternity index (TPI) were calculated for the 165 AI-SNPs in the Gannan Tibetan group, and the results are shown in **Supplementary Table 2.** Nine homozygous SNP loci showed the minimum values of H_{exp}, H_{obs}, PD, PE, PIC and TPI in the Gannan Tibetan group. Among the 165 AI-SNP loci, 71 AI-SNP loci showed H_{exp} values greater than 0.4 while 37 SNP loci showed H_{exp} less than 0.2. There were 79 SNP loci with PIC values greater than 0.3, whereas 25 loci showed relatively small PIC values (PIC <0.1). Except for nine homozygous SNP loci, rs3943253 locus displayed the largest PD value (0.6662), and the minimum values of PD were noted at rs12913832, rs17642714, and rs2042762 loci (PD = 0.0196). The largest PE value was 0.2957, which was observed at rs37369 locus, while the minimum PE value was 0.0001 at rs12913832, rs17642714, and rs2042762 loci. The combined PD (CPD) and combined PE (CPE) values for the 164 AI-SNP loci in the Gannan Tibetan group were 1-7.269 $\times 10^{-47}$ and 0.999999941, respectively (rs3811801 was excluded from the calculations due to LD).

3.4 Analyses of Genetic Ancestry of the Gannan Tibetan Group on the Basis of 164 AI-SNP Loci

Admixture predictions and population likelihood analyses for 101 Gannan Tibetans were first conducted using the plugin in Torrent Suite, and all individuals were predicted as East Asian individuals. When trying to predict the ancestral information of East Asian subgroups using this panel, the prediction results are not particularly ac-

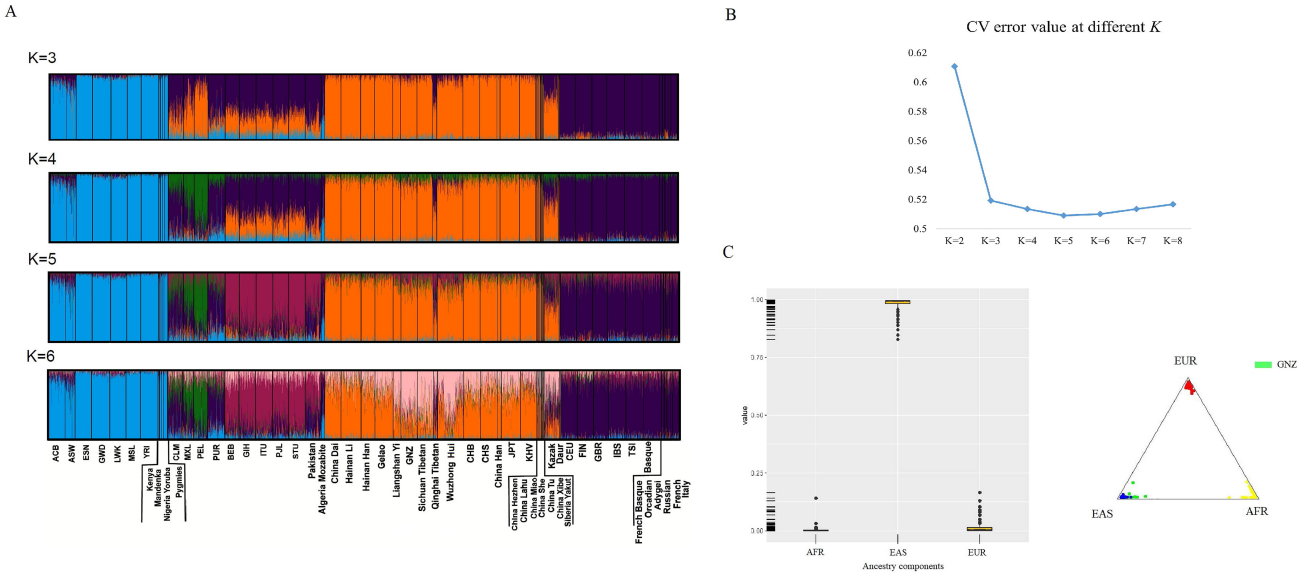

Fig. 2. Analyses of genetic ancestry for the Gannan Tibetan group (GNZ) on the basis of 164 AI-SNP loci. (A) Bayesian clustering analyses with predefined *K* values from 3 to 6. (B) Line chart of CV error values at *K* values from 2 to 8. (C) African, East Asian and European ancestral components in the Gannan Tibetan group.

curate (**Supplementary Fig. 3**). Genetic ancestry analyses for the Gannan Tibetan group were evaluated based on the 164 AI-SNP loci using both the ADMIXTURE (version 1.3, David H Alexander & John Novembre & Kenneth Lange, CA, USA) and STRUCTURE softwares (version 2.3.4, Pritchard & Stephens & Donnelly, UK), and the results are shown in Fig. 2. The Bayesian clustering analyses (*K* values are set to 2~6) and corresponding CV error values are shown in Fig. 2A,B, respectively. The smallest CV error value was found when $K = 5$ (Fig. 2B, CV error = 0.50892). Therefore, we adopted the $K = 5$ as the optimal *K* value, and all individuals were assigned to five different clusters, which could differentiate among African (a cyan-based ancestry cluster), American (a mixture of purple and green), South Asian (a brick-red-based ancestry cluster), East Asian (an orange-based ancestry cluster) and European populations (a purple-based ancestry cluster), as shown in Fig. 2A.

The Bayesian structure analyses indicated that the Gannan Tibetan group shared similar genetic ancestry structure with East Asian populations when *K* values were set to 3~5. Cluster analysis also showed that the Gannan Tibetans clustered tightly with the East Asian individuals (Fig. 2C). At $K = 6$, subcluster with pink-based ancestry structure was identified in Tibeto-Burman groups such as the Gannan Tibetan group, the Sichuan Tibetan group and the Liangshan Yi group. The Gannan Tibetan group displayed a mixture of pink and orange ancestry structure, which were different from Han populations in flat land. The studied Gannan Tibetan group shared similar ancestry structure with groups residing in its adjacent areas, such as the Sichuan Tibetan group and the Liangshan Yi group.

3.5 Population Genetic Relationship Analyses among the Gannan Tibetan group and Reference Populations

3.5.1 Heatmap and PCA plots among the Gannan Tibetan Group and Reference Populations Based on Ancestral Informative SNP Markers

To visualize the distributions of allele frequencies of 164 AI-SNPs (locus rs10954737 was excluded due to its unavailability in the 1000 Genomes Project) among the Gannan Tibetan group and reference populations, a heatmap was constructed, and the results are shown in Fig. 3. In the heatmap, the gray to green color scheme represented low to high allele frequencies of the 164 AI-SNP loci. Worldwide reference populations could be distinguished to four major clusters: (1) twelve populations from Europe clustered together into subbranch I; (2) most populations from Southern Asia gathered together and formed subbranch II; (3) most reference populations from China clustered together into subbranch III while Chinese Kazak group clustered between European and East Asian populations. At the same time, the studied Gannan Tibetan group clustered closely with Chinese groups, especially with the Sichuan Tibetan group; (4) eleven populations from Africa gathered into the subbranch IV. Meanwhile, the 164 AI-SNP loci could be divided into six subclusters (clusters A to F). SNP loci in clusters A and E displayed low allelic frequencies in the African populations, and these loci could distinguish the African populations from other continental reference populations; cluster D contained a set of SNP loci with low allelic frequencies in the East Asian populations and relatively large allelic frequencies in other continental reference populations; SNP loci in the subbranch F-1 showed higher allelic frequencies in African populations compared with other reference populations, and these loci could also

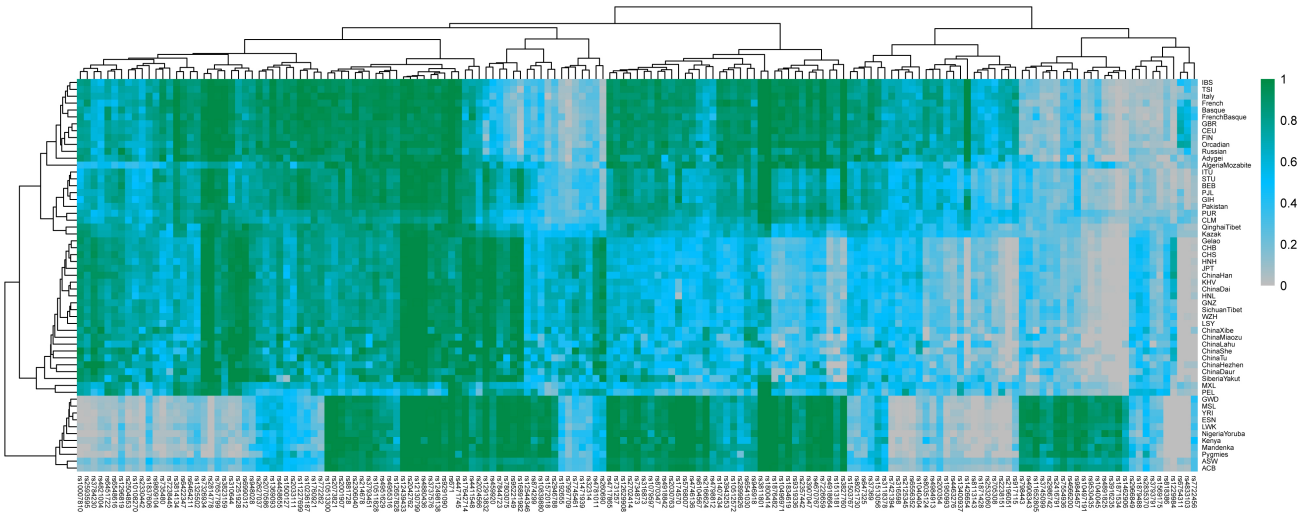

Fig. 3. Heatmap showing the minor allelic frequency distributions of 164 AI-SNPs among Gannan Tibetan group and 56 reference populations. The gray to green color scheme represented low to high allele frequency distributions of the 164 AI-SNP loci.

distinguish the African populations from other continental reference populations.

A series of PCA plots were generated to illustrate the genetic relations and differentiations between the Gannan Tibetan group and other reference populations, and the PCA results are shown in Fig. 4. The PCA plots at individual scale are shown in Fig. 4A (x-axis: PC1; y-axis: PC2) and Fig. 4B (x-axis: PC2; y-axis: PC3), respectively. And PC1 could distinguish the African individuals from non-African individuals, while PC2 could distinguish the European individuals and most East Asian individuals from other continental reference individuals. From Fig. 4A,B, individuals from worldwide populations gathered into four major clusters: Africa, Europe, East Asia and South Asia, labeled in red, blue, green and magenta, respectively. The studied Gannan Tibetan individuals clustered closely with the East Asian individuals, especially with the Chinese groups. A 3D-PCA plot was also generated using an online tool based on the PC1, PC2 and PC3, and the result is shown in Fig. 4C. The top three principal components could explain 44.19% total variance. Individuals from Africa, Europe, Oceania, South Asia, East Asia, Middle East and Americas were clustered together according to their geographical origins. Gannan Tibetan individuals clustered closely with East Asian individuals. A PCA plot at population scale was also generated (Fig. 4D). PC1 and PC2 could explain 77.2% total variance, and the studied Gannan Tibetan group was located closely with East Asian populations.

3.5.2 Phylogenetic Analyses and Genetic Differentiations between the Gannan Tibetan Group and the Reference Populations

Phylogenetic analyses were conducted using the NJ and ML methods to investigate the genetic relations between the Gannan Tibetan group and the reference populations, and the results are shown in Fig. 5. A NJ tree was constructed based on the pairwise D_A values, and the length of the connecting line in the bar plot around the NJ tree represented the pairwise F_{ST} values between Gannan Tibetan group and the reference populations (Fig. 5A). In the NJ tree, populations from Africa, Europe, South Asia and East Asia clustered together and formed four distinct subbranches accordingly. The studied Gannan Tibetan group was located in the East Asian subbranch, and clustered closely with the Sichuan Tibetan group and the Liangshan Yi group. In Fig. 5B, a rooted ML tree was constructed based on the genotypes of the 164 AI-SNPs. The obtained results were generally consistent with the NJ tree which demonstrated that the Gannan Tibetan group has tight phylogenetic connections with the East Asian populations, especially with the groups residing in its adjacent geographical areas.

We also calculated the pairwise F_{ST} values between Gannan Tibetan group and the continental reference populations, and the results are shown in the circular barchart around the NJ tree. Large F_{ST} values were observed between the Gannan Tibetan group and African populations. Moreover, the maximum F_{ST} value was observed between the Gannan Tibetan group and Yoruba in Ibadan, Nigeria ($F_{ST} = 0.4664$). Smaller F_{ST} values were found between the Gannan Tibetan group and East Asian populations, and the minimum F_{ST} value was found between the Gannan Tibetan group and the Sichuan Tibetan group.

4. Discussion

The read depths, strand balance, noise and other NGS quality control are important to evaluate the accuracy of NGS genotyping. Therefore, we first evaluated the sequencing metrics before the population genetic analyses. According to “Forensic sciences - Specifications for second generation sequencing-based DNA examination” (GA/T

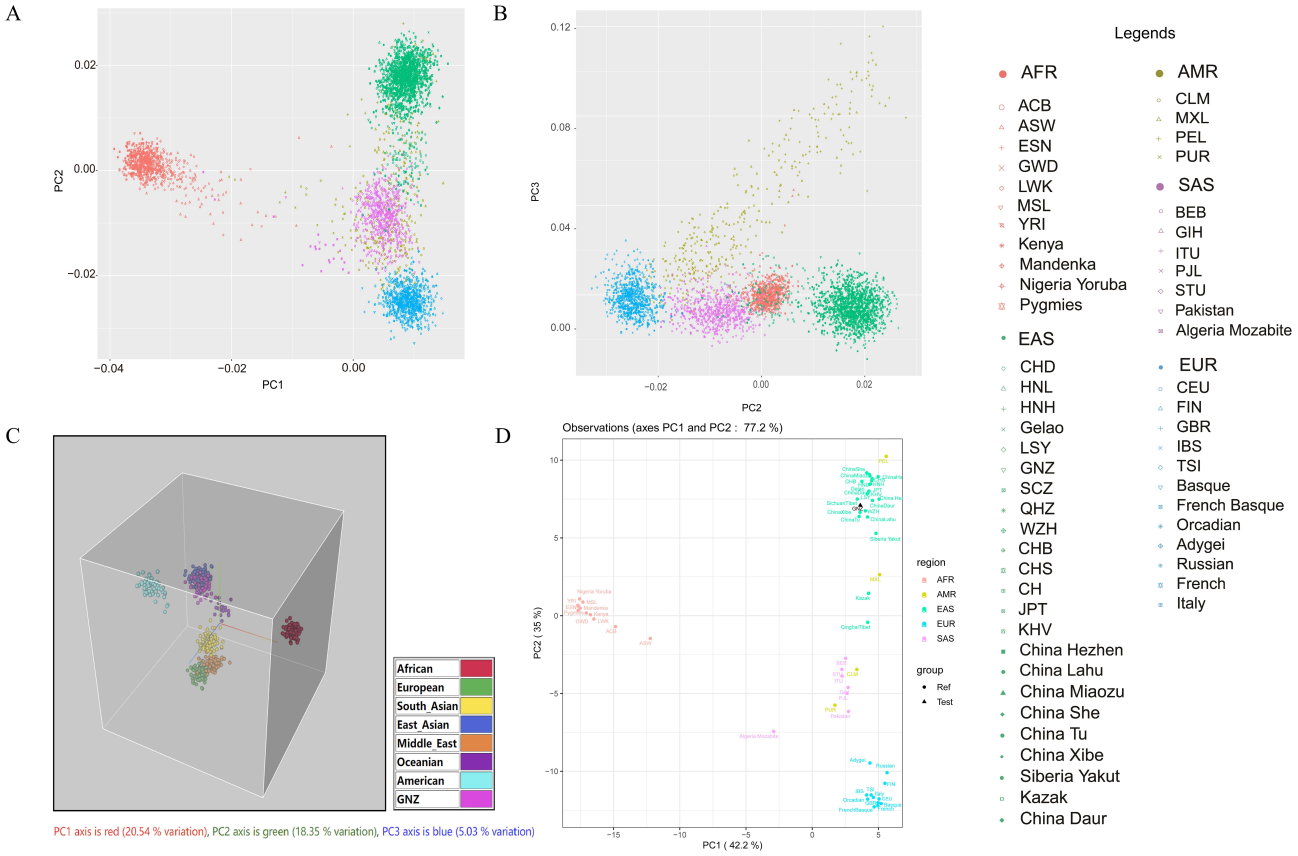

Fig. 4. Principal component analyses at different population scales. (A) PCA plot for 57 reference populations at individual scale based on PC1 and PC2. (B) PCA plot for 57 reference populations at individual scale based on PC2 and PC3. (C) A 3D-PCA plot was generated using an online tool based on the PC1, PC2 and PC3 to investigate the genetic relationships between Gannan Tibetan group and continental reference populations. (D) A PCA plot at population scale was generated based on PC1 and PC2.

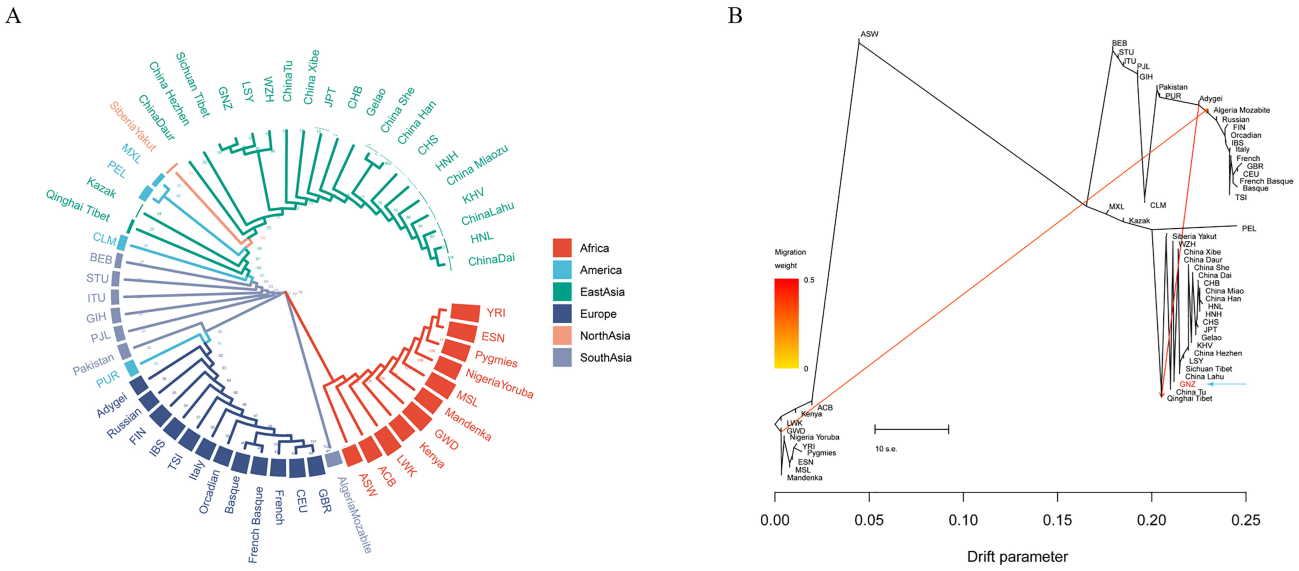

Fig. 5. Phylogenetic analyses were conducted using neighbor-joining (NJ) and maximum likelihood (ML) methods to investigate the genetic relations between the Gannan Tibetan group and the 56 reference populations. (A) A NJ tree was constructed based on the pairwise D_A values. The length of the connecting line in the barchart represented the pairwise F_{ST} values between the Gannan Tibetan group and the reference populations. (B) A rooted ML tree for 57 populations assumed that two migration events happened.

1693-2020), sequencing depth of SNP loci should be larger than 100×. In this study, the minimum depth was higher than the analysis threshold recommended by the guideline, the average sequencing depth was greater than 200×; and the strand balances of most amplicons were about 50%, indicating that the NGS genotyping results were reliable. In addition, relative low depth and notable inter-locus imbalance were observed at a few of AI-SNP loci in the Gannan Tibetan group. According to the validation studies and the population genetic researches based on this panel, notable inter-locus imbalance was also observed in the published articles, which might be related to the flanking sequence or the primers of the SNPs [23,29,42,43].

When we uploaded the genotype data to an online tool—Snipper (http://mathgene.usc.es/snipper/), the website reminded us to check the "A" allele at rs7722456 locus in the present study because only C and T alleles were observed at rs7722456 locus in their database. And then we searched the population genotype data of this SNP locus on 1000 Genomes Project Phase III, and also found that only C and T alleles were observed at rs7722456 locus. In the present study, allele 'A' at rs7722456 (chr 5:170775980) was found in 63 Gannan Tibetan individuals, and reads for the 'A' allele accounted for approximate 20% of the total reads. After sequencing the chr5:170775853 -170776519 region, we found that a poly-A region [5′-AATTACAGAT(C/T)AAAAAATACA-3′] was located immediately downstream of rs7722456 loci, which made it difficult for the Ion S5 XL system to identify the real number of poly-nucleotides in this region [42]. Certainly, raising the heterozygote balance threshold should be helpful to eliminate this deficiency [42,44].

Among the 165 AI-SNPs tested in this study, only one pair of rs1229984 and rs3811801 loci showed linkage disequilibrium, which was mainly caused by the inheritance of a haplotype encompassing these two SNP loci. Loci rs1229984 (chr 4:99318162, coding for the Arg48His substitution) and rs3811801 (chr 4:99323162) belong to the core haplotype of alcohol dehydrogenase family 1B gene (*ADH1B*), which was associated with the catalytic activity of the enzyme alcohol dehydrogenase [4]. Forensic parameters of the 165 AI-SNPs indicated that not all of the SNP loci showed high genetic polymorphisms in the Gannan Tibetan group. The cumulative random match probability of the 164 AI-SNPs was determined to be 7.269×10^{-47}, which was greater than that of the 27 autosomal STRs in the ForenSeq DNA Signature Prep Kit [45,46]. The CPD and CPE values of the 164 AI-SNPs in Gannan Tibetan group indicated that this panel could be an effective tool in the forensic individual identification and parentage testing of the Gannan Tibetan group.

The origin, migration and genetic background of Tibetan group have become the research hotspots in the field of population genetics, molecular anthropology, archaeology and linguistics [8,47–49]. In this study, we investigated the genetic ancestry structure of the Gannan Tibetan group on the basis of a set of AI-SNP loci. Results of Bayesian structure analyses indicated that the ancestral structure of the Gannan Tibetan group was dominated by the East Asian component, and the Gannan Tibetan group shared similar genetic ancestry structure with the Sichuan Tibetan group and the Liangshan Yi group.

To further analyze the genetic relationships among Gannan Tibetan group and reference populations, allelic frequency distributions, PCA, genetic distances and phylogenetic analyses were also conducted based on the shared AI-SNPs in the studied Gannan Tibetan group and continental reference populations. These analyses confirmed that the AI-SNP loci included in the Precision ID Ancestry Panel could be a valuable tool for the ancestral predictions of African, European, South Asian and East Asian populations. When trying to predict the ancestral information of East Asian subpopulations using this panel, the prediction results are not particularly accurate. An extended set of SNPs would be required to distinguish within East Asian subpopulations.

Results of genetic affinity analyses indicated that the Gannan Tibetan group had closer genetic affinities with East Asian populations, especially tighter genetic ties with those groups residing in its neighboring geographical regions like the Sichuan Tibetan group and the Liangshan Yi group. Many researchers estimated that the Sino-Tibetan groups shared a common ancestor who originated from the upper and middle reaches of Yellow River. With the spread of agriculture, ancient Han Chinese and Tibetan group derived from the shared ancestors about 5900 years ago [49–51]. Linguistic study and archaeological evidence also inferred that Tibeto-Burman populations consecutively migrated southward to be dispersed along the Tibetan-Yi ethnic corridor into the western Sichuan and western Yunnan provinces [49,52]. Population genetic evidences also demonstrated that the studied Gannan Tibetan group had closer genetic relationships with East Asian populations, which might be due to the extensive gene exchanges among Tibetan, Han Chinese and other groups residing around the Tibetan Plateau [2,29,47]. Besides, the plateau geographical environment also shaped the unique genetic characteristics of the Tibetan group. Frequencies of SNP haplotypes associated with high-altitude hypoxia adaptation were higher than those in populations residing in the plain region [53–55]. Therefore, these SNP loci mentioned above might be regarded as AIM-SNP markers to distinguish the Tibetan group and Han populations. Population genetic studies based on Y-SNP haplogroups showed that the Tibetan group was dominated by D-M174 haplogroup, followed by O-M175 haplogroup, which had relative high frequencies in Han population. The mtDNA haplogroups (M9a'b, G, D and F) of East Asian origin had high frequencies in the Tibetan group, which also indicated the close genetic relationships among the Tibetan group and East Asian popula-

tions [56]. In the future, more types of genetic markers will be used for population genetic analysis, which will allow us to have a deeper understanding of the population genetic sub-structures of the Tibetan groups in Chinese different regions.

5. Conclusions

In this study, the NGS performance and forensic statistical parameters of the 165 AI-SNP loci in the Gannan Tibetan group were evaluated. The 165 AI-SNP loci showed the varying degrees of genetic polymorphisms in the Gannan Tibetan group, and the combined use of these loci could be an effective tool in the forensic individual identification and parentage testing of the Gannan Tibetan group. The ancestral information component of the Gannan Tibetan group was dominated by the East Asian component. Population genetic analyses indicated that the Gannan Tibetan group had close genetic affinities with East Asian populations compared with other intercontinental reference populations, especially tighter genetic relationships with the groups residing in its neighboring geographical regions.

Abbreviations

AIM, ancestry informative marker; CPD, combined power of discrimination; CPE, combined probability of exclusion; D_A, *Nei's* genetic distances; DIP, deletion-insertion polymorphism; F_{ST}, fixation index; H_{exp}, expected heterozygosity; H_{obs}, observed heterozygosity; HWE, Hardy-Weinberg equilibrium; LD, linkage disequilibrium; ML, maximum likelihood; MP, matching probability; NGS, next generation sequencing; PCA, principal component analysis; PD, power of discrimination; PE, probability of exclusion; PIC, polymorphism information content; SNP, single nucleotide polymorphisms; TPI, typical paternity index.

Availability of Data and Materials

The raw genotype data used and analyzed during the current study are available from the corresponding author on reasonable request.

Author Contributions

BZ designed this research and were responsible for all the processes of this research. WC and MC conducted the experiment and analyzed the raw data. WC wrote this manuscript. HY collected the samples. QY and LC provided technical support for the experiments. LC, LL and XB assisted the experiment. BZ, MC, LL, LC and XB revised this manuscript. All authors have read and agreed to submit the manuscript.

Ethics Approval and Consent to Participate

This study has been approved by the ethics committee of Xi'an Jiaotong University Health Science Center and Southern Medical University (Ethical Approval Number: 2019-1039). This research was conducted in accordance with the ethical principle for medical research involving human subjects recommended by the World Medical Association Declaration of Helsinki. Sample collections and sequencing experiments presented herein were performed in strict compliance with the ethical regulation. Every volunteer signed a written informed consent before sample collection.

Acknowledgment

We are grateful to all of the volunteers for their kind donations of samples.

Funding

This research was funded by the National Natural Science Foundation of China (NSFC), grant numbers 81930055 and 31760309.

Conflict of Interest

Qing Yang is the employee of the Thermo Fisher Scientific (China) Inc. Other authors declare no conflict of interest.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.31083/j.fbl2806114.

References

[1] Xu S, Li S, Yang Y, Tan J, Lou H, Jin W, *et al*. A genome-wide search for signals of high-altitude adaptation in Tibetans. Molecular Biology and Evolution. 2011; 28: 1003–1011.

[2] Jin XY, Shen CM, Chen C, Guo YX, Cui W, Wang YJ, *et al*. Ancestry informative DIP loci for dissecting genetic structure and ancestry proportions of Qinghai Tibetan and Tibet Tibetan groups. Molecular Biology Reports. 2020; 47: 1079–1087.

[3] Office of the Leading Group of the State Council for the Seventh National Population Census. China Population Census Yearbook-2020. China Statistics Press: Beijing. 2020.

[4] Lu Y, Kang L, Hu K, Wang C, Sun X, Chen F, *et al*. High diversity and no significant selection signal of human ADH1B gene in Tibet. Investigative Genetics. 2012; 3: 23.

[5] People's Government of Gannan Tibetan Autonomous Prefecture. Introductions to Gannan Tibetan Autonomous Prefecture. 2021. Available at http://www.gnzrmzf.gov.cn/zjgn/gngk/zrdl.htm (Accessed: 11 October 2022).

[6] Wang M, Wang Z, He G, Wang S, Zou X, Liu J, *et al*. Whole mitochondrial genome analysis of highland Tibetan ethnicity using massively parallel sequencing. Forensic Science International. Genetics. 2020; 44: 102197.

[7] Liu Y, Jin X, Mei S, Xu H, Zhao C, Lan Q, *et al*. Insights into the genetic characteristics and population structures of Chinese two Tibetan groups using 35 insertion/deletion polymorphic loci. Molecular Genetics and Genomics. 2020; 295: 957–968.

[8] Li G, Lin Y, Lan S, Zou J, Li S, Song F, *et al*. Tibetan Y-STR trait in the eleven regions of the Qinghai-Tibet Plateau. International Journal of Legal Medicine. 2021; 135: 1793–1795.

[9] Phillips C, Santos C, Fondevila M, Carracedo Á, Lareu MV. In-

ference of Ancestry in Forensic Analysis I: Autosomal Ancestry-Informative Marker Sets. Methods in Molecular Biology. 2016; 1420: 233–253.
[10] Halder I, Shriver M, Thomas M, Fernandez JR, Frudakis T. A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. Human Mutation. 2008; 29: 648–658.
[11] Zhang X, Shen C, Jin X, Guo Y, Xie T, Zhu B. Developmental validations of a self-developed 39 AIM-InDel panel and its forensic efficiency evaluations in the Shaanxi Han population. International Journal of Legal Medicine. 2021; 135: 1359–1367.
[12] Bulbul O, Pakstis AJ, Soundararajan U, Gurkan C, Brissenden JE, Roscoe JM, *et al*. Ancestry inference of 96 population samples using microhaplotypes. International Journal of Legal Medicine. 2018; 132: 703–711.
[13] Phillips C, Prieto L, Fondevila M, Salas A, Gómez-Tato A, Alvarez-Dios J, *et al*. Ancestry analysis in the 11-M Madrid bomb attack investigation. PLoS ONE. 2009; 4: e6583.
[14] Suarez-Pajes E, Díaz-de Usera A, Marcelino-Rodríguez I, Guillen-Guio B, Flores C. Genetic Ancestry Inference and Its Application for the Genetic Mapping of Human Diseases. International Journal of Molecular Sciences. 2021; 22: 6962.
[15] Harris AM, DeGiorgio M. Admixture and Ancestry Inference from Ancient and Modern Samples through Measures of Population Genetic Drift. Human Biology. 2017; 89: 21–46.
[16] Sun Q, Jiang L, Zhang G, Liu J, Zhao L, Zhao W, *et al*. Twenty-seven continental ancestry-informative SNP analysis of bone remains to resolve a forensic case. Forensic Sciences Research. 2017; 4: 364–366.
[17] Shi CM, Liu Q, Zhao S, Chen H. Ancestry informative SNP panels for discriminating the major East Asian populations: Han Chinese, Japanese and Korean. Annals of Human Genetics. 2019; 83: 348–354.
[18] Jung JY, Kang PW, Kim E, Chacon D, Beck D, McNevin D. Ancestry informative markers (AIMs) for Korean and other East Asian and South East Asian populations. International Journal of Legal Medicine. 2019; 133: 1711–1719.
[19] Cao Y, Zhu Q, Huang Y, Li X, Wei Y, Wang H, *et al*. An efficient ancestry informative SNPs panel for further discriminating East Asian populations. Electrophoresis. 2022; 43: 1774–1783.
[20] Gu JQ, Zhao H, Guo XY, Sun HY, Xu JY, Wei YL. A high-performance SNP panel developed by machine-learning approaches for characterizing genetic differences of Southern and Northern Han Chinese, Korean, and Japanese individuals. Electrophoresis. 2022; 43: 1183–1192.
[21] Chen L, Zhou Z, Zhang Y, Xu H, Wang S. EASplex: A panel of 308 AISNPs for East Asian ancestry inference using next generation sequencing. Forensic Science International. Genetics. 2022; 60: 102739.
[22] Al-Asfi M, McNevin D, Mehta B, Power D, Gahan ME, Daniel R. Assessment of the Precision ID Ancestry panel. International Journal of Legal Medicine. 2018; 132: 1581–1594.
[23] He G, Liu J, Wang M, Zou X, Ming T, Zhu S, *et al*. Massively parallel sequencing of 165 ancestry-informative SNPs and forensic biogeographical ancestry inference in three southern Chinese Sinitic/Tai-Kadai populations. Forensic Science International. Genetics. 2021; 52: 102475.
[24] Shan MA, Meyer OS, Refn M, Morling N, Andersen JD, Børsting C. Analysis of Skin Pigmentation and Genetic Ancestry in Three Subpopulations from Pakistan: Punjabi, Pashtun, and Baloch. Genes. 2021; 12: 733.
[25] QIAGEN. Introductions to QIAamp DNA Investigator Kit. 2022. Available at https://www.qiagen.com/cn/products/human-id-and-forensics/investigator-solutions/qiaamp-dna-investigator-kit (Accessed: 11 October 2022).
[26] Mardis E, McCombie WR. Library Quantification: Fluorometric Quantitation of Double-Stranded or Single-Stranded DNA Samples Using the Qubit System. Cold Spring Harbor Protocols. 2017; 2017: pdb.prot094730.
[27] 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, *et al*. A map of human genome variation from population-scale sequencing. Nature. 2010; 467: 1061–1073.
[28] Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, *et al*. A human genome diversity cell line panel. Science. 2002; 296: 261–262.
[29] Wang Z, He G, Luo T, Zhao X, Liu J, Wang M, *et al*. Massively parallel sequencing of 165 ancestry informative SNPs in two Chinese Tibetan-Burmese minority ethnicities. Forensic Science International. Genetics. 2018; 34: 141–147.
[30] He G, Wang Z, Wang M, Luo T, Liu J, Zhou Y, *et al*. Forensic ancestry analysis in two Chinese minority populations using massively parallel sequencing of 165 ancestry-informative SNPs. Electrophoresis. 2018; 39: 2732–2742.
[31] Xie T, Shen C, Liu C, Fang Y, Guo Y, Lan Q, *et al*. Ancestry inference and admixture component estimations of Chinese Kazak group based on 165 AIM-SNPs via NGS platform. Journal of Human Genetics. 2020; 65: 461–468.
[32] García O, Ajuriagerra JA, Alday A, Alonso S, Pérez JA, Soto A, *et al*. Frequencies of the precision ID ancestry panel markers in Basques using the Ion Torrent PGM™ platform. Forensic Science International. Genetics. 2017; 31: e1–e4.
[33] Gouy A, Zieger M. STRAF-A convenient online tool for STR data evaluation in forensic genetics. Forensic Science International. Genetics. 2017; 30: 148–151.
[34] Excoffier L, Laval G, Schneider S. Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online. 2007; 1: 47–50.
[35] Tatsuya Ota. DISPAN: Genetic Distance and Phylogenetic Analysis [Master's thesis]. Pennsylvania State University. 1993.
[36] Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Molecular Biology and Evolution. 2016; 33: 1870–1874.
[37] Yu GC, Smith DK, Zhu HC, Guan Y, Lam TTY. GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution. 2017; 8: 28–36.
[38] Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genetics. 2012; 8: e1002967.
[39] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics. 2006; 38: 904–909.
[40] Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Research. 2009; 19: 1655–1664.
[41] Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology. 2005; 14: 2611–2620.
[42] Pereira V, Mogensen HS, Børsting C, Morling N. Evaluation of the Precision ID Ancestry Panel for crime case work: A SNP typing assay developed for typing of 165 ancestral informative markers. Forensic Science International. Genetics. 2017; 28: 138–145.
[43] Guo F, Zhou Y, Song H, Zhao J, Shen H, Zhao B, *et al*. Next generation sequencing of SNPs using the HID-Ion AmpliSeq™ Identity Panel on the Ion Torrent PGM™ platform. Forensic Science International. Genetics. 2016; 25: 73–84.
[44] Børsting C, Morling N. Next generation sequencing and its applications in forensic genetics. Forensic Science International. Genetics. 2015; 18: 78–89.

[45] Li H, Zhang C, Song G, Ma K, Cao Y, Zhao X, *et al*. Concordance and characterization of massively parallel sequencing at 58 STRs in a Tibetan population. Molecular Genetics & Genomic Medicine. 2021; 9: e1626.
[46] Chen C, Jin X, Zhang X, Zhang W, Guo Y, Tao R, *et al*. Comprehensive Insights Into Forensic Features and Genetic Background of Chinese Northwest Hui Group Using Six Distinct Categories of 231 Molecular Markers. Frontiers in Genetics. 2021; 12: 705753.
[47] Yao HB, Wang CC, Wang J, Tao X, Shang L, Wen SQ, *et al*. Genetic structure of Tibetan populations in Gansu revealed by forensic STR loci. Scientific Reports. 2017; 7: 41195.
[48] Ding M, Wang T, Ko AMS, Chen H, Wang H, Dong G, *et al*. Ancient mitogenomes show plateau populations from last 5200 years partially contributed to present-day Tibetans. Proceedings. Biological Sciences. 2020; 287: 20192968.
[49] Zhang M, Yan S, Pan W, Jin L. Phylogenetic evidence for Sino-Tibetan origin in northern China in the Late Neolithic. Nature. 2019; 569: 112–115.
[50] Wang CC, Yeh HY, Popov AN, Zhang HQ, Matsumura H, Sirak K, *et al*. Genomic insights into the formation of human populations in East Asia. Nature. 2021; 591: 413–419.
[51] Wang LX, Lu Y, Zhang C, Wei LH, Yan S, Huang YZ, *et al*. Reconstruction of Y-chromosome phylogeny reveals two neolithic expansions of Tibeto-Burman populations. Molecular Genetics and Genomics: MGG. 2018; 293: 1293–1300.
[52] Li KS. Collections of Yunnan Archaeology (Yunnan Kaoguxue Lunji, in Chinese). Yunnan People's Publishing House: Kunming, Yunnan. 1998.
[53] Xiang K, Ouzhuluobu, Peng Y, Yang Z, Zhang X, Cui C, *et al*. Identification of a Tibetan-specific mutation in the hypoxic gene EGLN1 and its contribution to high-altitude adaptation. Molecular Biology and Evolution. 2013; 30: 1889–1898.
[54] Zhang X, Witt KE, Bañuelos MM, Ko A, Yuan K, Xu S, *et al*. The history and evolution of the Denisovan-*EPAS1* haplotype in Tibetans. Proceedings of the National Academy of Sciences of the United States of America. 2021; 118: e2020803118.
[55] Chen Y, Jiang C, Luo Y, Liu F, Gao Y. An EPAS1 haplotype is associated with high altitude polycythemia in male Han Chinese at the Qinghai-Tibetan plateau. Wilderness & Environmental Medicine. 2014; 25: 392–400.
[56] Wang XJ, Qian EF, Li Y, Song ZY, Zhao H, Xie HX, *et al*. A genetic sub-structure study of the Tibetan population in Southwest China (in Chinese). Yi Chuan. 2020; 42: 565–576.