BMC Genomics

RESEARCH Open Access

Paternal genetic structure and Y-chromosomal haplogroup prediction in the Tujia and Bai ethnic groups of Guizhou, Western China

Wei Wang[1*†], Xiaoyuan Zhen[2†], Han Zhang[3†], Juncheng Zhao[3], Nan An[3], Jinfeng Ding[1], Peiyan Qi[3] and Meiqing Yang[3*]

Abstract

Background Y chromosome genetic markers, with strict paternal inheritance and lack of recombination, are particularly valuable tools for tracing male lineages. They complement autosomal analyses in forensic applications and anthropological inference by increasing resolution for patrilineal structure.

Methods We genotyped 382 unrelated male individuals from two Tibeto-Burman-speaking populations in Guizhou (Tujia, $n=220$; Bai, $n=162$) using the Goldeneye DNA Identification System Y Plus kit comprising 44 Y-markers. We calculated haplotype-level forensic indices and assessed inter-population structure via Rst-based multidimensional scaling (MDS) and a neighbor-joining (NJ) tree, based on genetic distances with 47 reference groups. Y-chromosomal haplogroups were predicted from Y-STR profiles to characterize paternal lineages.

Results In the Tujia population, 338 alleles and 219 haplotypes were detected, with allelic frequencies ranging from 0.0045 to 0.9364. The haplotype diversity (HD), haplotype match probability (HMP), and discrimination capacity (DC) were 0.9999, 0.0046, and 0.9955, respectively. In the Bai population, 309 alleles and 141 haplotypes were detected, with allelic frequencies ranging from 0.0062 to 0.9691, with HD = 0.9979, HMP = 0.0083, and DC = 0.8704. Population genetic analysis revealed that the Guizhou Tujia and Bai groups share closer genetic affinity with Southern Han than with Northern Han and cluster with Tibeto-Burman-speaking groups, including the Sichuan and Guizhou Yi populations. Similarly, the Y-STR haplogroup prediction results revealed a multilayered paternal structure dominated by haplogroup O2a2, accompanied by contributions from indigenous East Asian lineages and minor inputs from West Eurasia and other regions.

Conclusions Our study provides valuable Y-STR data and forensic parameters for Tibeto-Burman-speaking ethnic groups in China, as well as population genetics evidence in patrilineal history. The Tujia and Bai of Guizhou exhibit a complex paternal genetic structure, offering insights into the demographic dynamics of Southwest China. The 43

†Wei Wang, Xiaoyuan Zhen and Han Zhang contributed equally to this work.

*Correspondence:
Wei Wang
xzww1993@126.com
Meiqing Yang
ymq918@126.com

Full list of author information is available at the end of the article

Y-marker system exhibits high polymorphism and strong discriminatory power, supporting its utility as a powerful supplementary tool for forensic investigations, particularly for male lineage inference and suspect screening.

Keywords Y-STR, Paternal genetic structure, Guizhou, Southwest China

Introduction

Y-chromosome short tandem repeats (Y-STRs) have played an invaluable role in forensic science due to their unique mode of paternal inheritance. They are widely applied to resolve paternity disputes, identify male victims in disaster and missing persons cases, and establish male involvement in sexual assault and familial investigations [1, 2]. The non-recombining region of the Y chromosome makes Y-STR haplotypes essential for inferring biogeographical ancestry and population origins, facilitating their broad application in forensic genetics, population history, and population genetics [3–5]. The Y-STR Haplotype Reference Database (YHRD, https://yhrd.org) is an open-access repository that compiles Y-STR and Y-SNP haplotypes from diverse global populations. It estimates haplotype frequencies within defined subsets and accounts for population substructure based on geographical, regional, and ethnolinguistic factors [6]. The development of commercial Y-STR kits designed for forensic research, such as the PowerPlex Y23 [7] and Goldeneye Y-Plus kit (Goldeneye® DNA Identification System Y-Plus Kit; http://www.peoplespot.com.cn /products_details/7.html), has facilitated their wide ap plication in population genetics studies and contributed to data enrichment across different populations [8–10]. Nonetheless, population data for certain geographical subgroups, particularly minority ethnic groups, remain limited. Additionally, forensic practices necessitate the further enrichment of databases with new Y-STR loci data.

Guizhou province, located in the southwest of China, has a highly diverse population in terms of both ethnicity and language. According to the data from the Seventh National Population Census in 2020, the Tujia ethnic group, one of the oldest ethnicities in Guizhou, numbers over 1.6 million, accounting for 17.7% of the national Tujia population and approximately 4.4% of the total provincial population, with most residing in the northeastern part of the province. In addition, since the 1980s, the Bai ethnic group has also been recognized as an indigenous ethnicity in Guizhou. The Bai population in the province reaches 0.21 million, representing 10.3% of the national Bai population and 0.6% of the provincial population, with Yunnan and Guizhou serving as the main distribution areas of this group. The Tujia and Bai peoples both speak Tibeto-Burman language, which forms a significant subbranch of the Sino-Tibetan language family, and those spoken in southwest China have undergone substantial influence from Chinese [11, 12]. Previous genetic data on Tibeto-Burman-speaking populations from the Tibetan-Yi Corridor (TYC), including Tibetans, Yi people from Gansu, Sichuan, and Yunnan, have been reported [12–15]. Moreover, these studies have highlighted the pivotal role of the TYC in the southward expansion of Tibeto-Burman populations, with certain branches migrating eastward to the Guizhou region [16]. However, the ethnic origins and genetic structure of Tibeto-Burman-speaking groups in Guizhou (TBG) remain largely unexplored.

In our study, we conducted a comprehensive analysis of the genetic distribution patterns across 43 Y-markers among male individuals from Tibeto-Burman-speaking populations, including 220 Tujia and 162 Bai individuals in Guizhou province. Furthermore, we examined the genetic similarity between the studied populations and 47 additional reference populations. This research significantly enriches the existing genetic database of Y-markers, providing valuable insights into the paternal genetic structure of Tibeto-Burman-speaking populations residing in Guizhou.

Materials & methods

Sample information

We recruited 382 unrelated healthy male individuals from Guizhou province, China, including 220 Tujia and 162 Bai participants. All individuals reported continuous residence of their families in Guizhou for at least three generations. Peripheral bloodstain samples were collected after written informed consent was obtained from each participant. The study protocol was approved by the Ethics Commission of Guizhou Medical University, Approval No. 2021063.

PCR amplification and capillary electrophoresis

PCR amplification was performed on a GeneAmp 9700 thermal cycler (Applied Biosystems, Foster City, CA, USA), using the Goldeneye® Y-Plus kit, following the manufacturer's instructions. The PCR reaction was prepared with a total volume of 10μL, comprising 2μL reaction mix, 2μL primers, 1μL Taq DNA polymerase, 1μL template DNA and 4μL sterile distilled water (sdH_2O). Thermal cycling conditions were identical to those described previously [17]. Amplified products were detected on an ABI 3500 Genetic Analyzer (Applied Biosystems). The sample 9948 was used as a positive control, and sdH_2O was used as the negative control in each run. Finally, genotypes were analyzed using GeneMapper ID-X v1.3 software (Thermo Fisher Scientific) to conduct the allele nomenclature by comparison with the

corresponding allelic ladder. Notably, a substantial proportion of genotypes at locus DYS522 failed during laboratory processing, and the locus was therefore excluded from downstream analyses.

Statistical analysis

Analyses were performed on 43 Y-markers (DYS522 excluded), comprising 40 Y-STRs and 3 Y-InDels (rs199815934, rs771783753, rs759551978). The 40 Y-STR markers consist of 32 single-copy loci (DYS19, DYS388, DYS389 I, DYS389 II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, Y-GATA H4, DYS449, DYS456, DYS458, DYS460, DYS481, DYS518, DYS533, DYS570, DYS576, DYS627, DYS635, DYS549, DYS444, DYS643, DYS447, DYS557, DYS596, DYS593, DYS645), and four multiple-copy loci (DYS385, DYF387S1, DYS527, DYF404S1). Among these, DYF387S1, DYF404S1, DYS449, DYS570, DYS576, and DYS627 were identified as rapidly mutating (RM) Y-STRs.

Allele and haplotype frequencies were computed by direct counting. Gene diversity (GD) and haplotype diversity (HD) were calculated based on the formula: $GD/HD = n(1-?P_i^2)/(n-1)$, where $P_i$is the frequency of the i^{th} allele or haplotype and n is the sample size [18]. The haplotype match probability (HMP) and discrimination capacity (DC) were calculated as $HMP = ?P_i^2 (P_i$ is the frequency of the i^{th} haplotype) [19], and $DC = N_{diff}/N (N_{diff}$ represents the count of distinct haplotypes and N refers to the sample size) [20].

For inter-population analyses, we used a subset of 27 Y-STR loci overlapping the Yfiler Plus kit (Applied Biosystems, Thermo Fisher Scientific), with reference data sourced from previously published studies [6]. The reference groups included 10 Northern Han populations [21–26], and 8 Southern Han populations [27–32], as well as 29 other ethnic minorities [8, 33–42]. Pairwise population genetic distances (Rst) and their significance (p values) were calculated in Arlequin v3.5 based on 10,000 random permutations. DYS385 was treated as two markers (DYS385a and DYS385b). Very rare multi-allelic profiles indicating additional copies were excluded from Rst analyses (effective sample sizes: Tujia n = 216, Bai n = 158). A neighbor-joining (NJ) tree based on Rst genetic distances among populations was constructed using MEGA7. The tree was rooted using the midpoint method [43], and multidimensional scaling (MDS) was performed using the SPSS v26.0 based on pairwise Rst distances.

Y-chromosomal haplogroups were inferred using the NevGen Y-DNA Haplogroup Predictor (http://www.nevgen.org), a publicly accessible web-based tool. We extracted the 30 Y-STR loci from our previously genotyped 44-marker Goldeneye dataset that are supported by NevGen and arranged them according to the required input order. Haplogroup probability estimates were then obtained for each haplotype. Haplogroup assignments with a predicted probability ≥ 0.80 were considered reliable.

Results

Allele frequencies and forensic characteristics of 43 Y-markers

All Y-STRs were typed using the kit's allele ladder, and all observed alleles matched the corresponding ladder bins (Table S1). At the DYS522 locus, 31.8% of Tujia individuals and 37.0% of Bai individuals showed no detectable allele, and this marker was therefore excluded from subsequent analyses. Allele distribution and GD values for 40 Y-STRs and 3 Y-InDels in the Guizhou Tujia and Bai populations were illustrated in Fig. 1 and detailed in Tables S1 and S2. Across 220 Tujia males, a total of 332 distinct alleles were detected across Y-STR loci, ranging from 2 (DYS645) to 37 (DYS385) per locus, with allele frequencies ranging from 0.0045 to 0.9364 (Table S2). The DYF387S1 locus exhibited the highest GD value (0.9530), whereas DYS645 showed the lowest (0.1197). Seven Y-markers (DYS391, DYS437, DYS438, DYS645, rs199815934, rs759551978, and rs771783753), including four Y-STRs and three Y-InDels, exhibited comparatively low GD values (< 0.5). Notably, the four multi-copy loci DYS385, DYS527, DYF404S1, and DYF387S1 all exhibited high gene diversity (GD > 0.9), with GDs of 0.9405, 0.9530, 0.9171, and 0.9329, respectively. Biallelic variants were detected at the single copy loci DYS570 (15,18/18,19) and DYS576 (17,20/18,19), and triallelic variants were observed at multi-copy loci DYF387S1 (36,37,38) and DYS527 (21,22,23).

Among the 162 Bai males, 309 alleles were identified, with allele frequencies ranging from 0.0062 to 0.9691 (Table S3). The highest GD value was observed at DYS385 (0.9383), and the lowest was at DYS645 (0.0607). There were five Y-markers (DYS391, DYS645, rs199815934, rs771783753, and rs759551978), including two Y-STRs and three Y-InDel loci, which displayed relatively low GD values (< 0.5), whereas three loci (DYS385, DYF387S1, and DYS527) exceeded 0.9. Triallelic patterns were detected at DYF387S1 (34,35,40/36,37,38/37,39,41) and DYS527 (20,21,25/21,22,23). This finding was consistent with previous reports [17, 42, 44] in which DYF387S1 frequently exhibited a triallelic pattern (36, 37, 38).

Comparison of forensic efficacy of different Y-marker sets in the Guizhou Tujia and Bai populations

Comparative analysis was done by restricting the number of loci in the Goldeneye Y-Plus dataset to those included in other commercial Y-STR kits (Minimal, PowerPlex Y, Yfiler, PowerPlex Y23, Argus Y-28, Yfiler Plus, AGCU Y37, and Yfiler Platinum), without performing additional

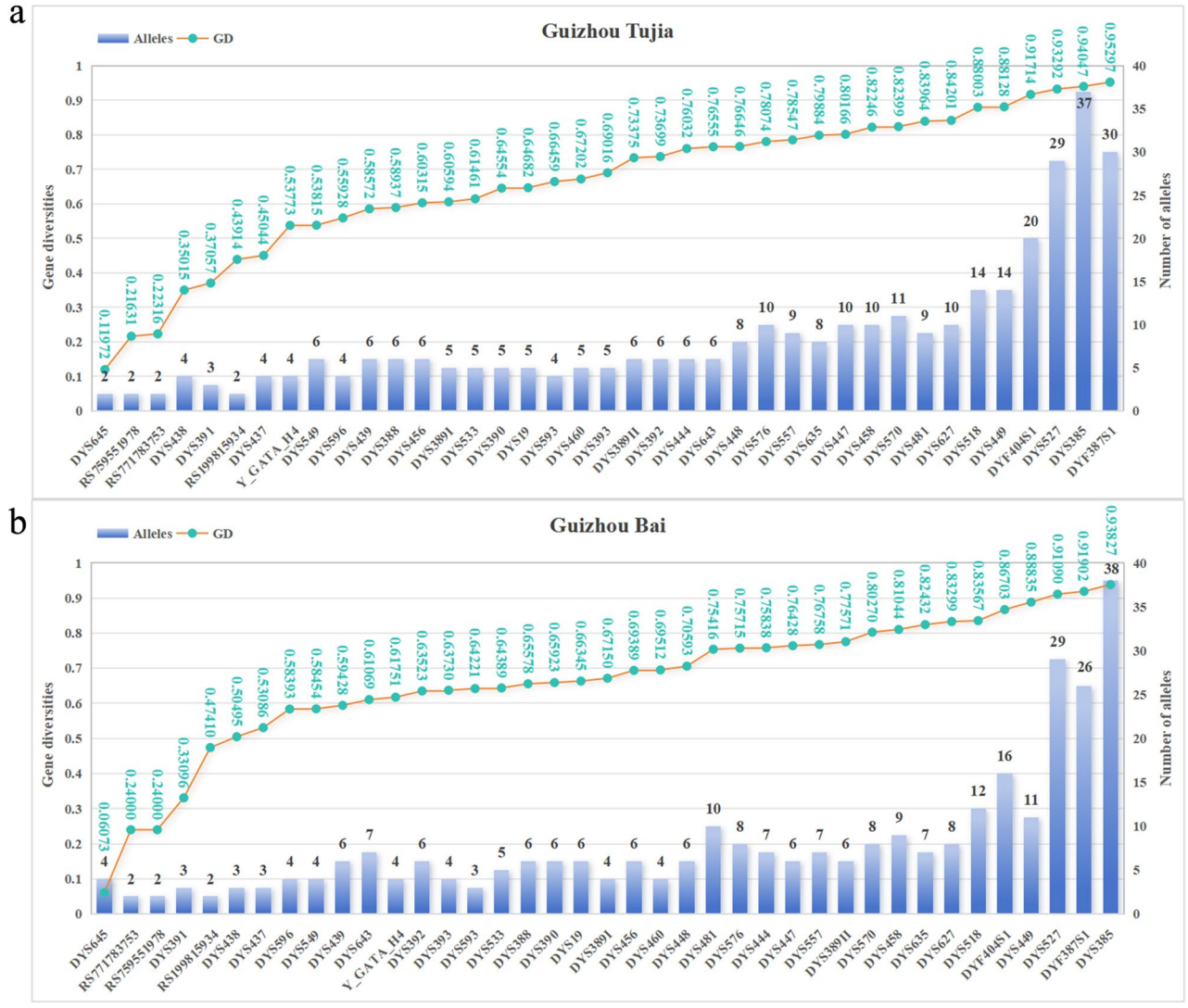

Fig. 1 The numbers of alleles and gene diversities of Y-markers in Guizhou Tujia and Bai populations. The orange line represents gene diversity (GD) values, and the blue bars represent allele counts. The right Y-axis indicates the number of alleles, and the left Y-axis indicates GD values

genotyping. Figure 2 summarizes haplotype numbers and forensic parameters across varying Y-STR panels. Detailed values for DC, HD, HMP, and haplotype numbers are provided in Table S4. Across progressively larger Y-STR panels, increasing locus count yielded more haplotypes and higher DC, with a concomitant decrease in HMP. Beyond the Argus Y-28, additional gains were marginal. HD likewise increased with panel size, but the increments were minimal.

In the Guizhou Tujia population, the 40 Y-STRs of the Goldeneye Y-Plus panel yielded higher DC (0.9954) and HD (0.999958) than other commercial sets evaluated (Table S4a). Among 220 individuals, 219 haplotypes were observed, of which 218 were unique (99.5%). In the Bai cohort, a total of 141 distinct haplotypes were identified among 162 males with the 40 Y-STR panel, with 90.07% being unique (Table S4b). The forensic parameters were HMP = 0.00831, HD = 0.99785, and DC = 0.8704. The DC increased from 0.6296 with the minimal 9-locus Y-STR set to 0.8704 with the 40-locus panel used in this study. Likewise, HD rose from 0.9817 (9 loci) to 0.99785 (40 loci).

Genetic relationships between the Guizhou Tujia and Bai populations and other reference populations

We assessed genetic affinities of two Tibeto-Burman-speaking groups from Guizhou (Tujia and Bai) relative to 18 Han populations and 29 ethnic minority populations across China by computing pairwise Rst distances (Tables S5 and S6) and visualizing the results with a clustered heatmap (Fig. 3).

Overall, the Tujia and Bai ethnic groups in Guizhou show a smaller Rst genetic distance to Southern Han populations, particularly those from Anhui and Guangxi

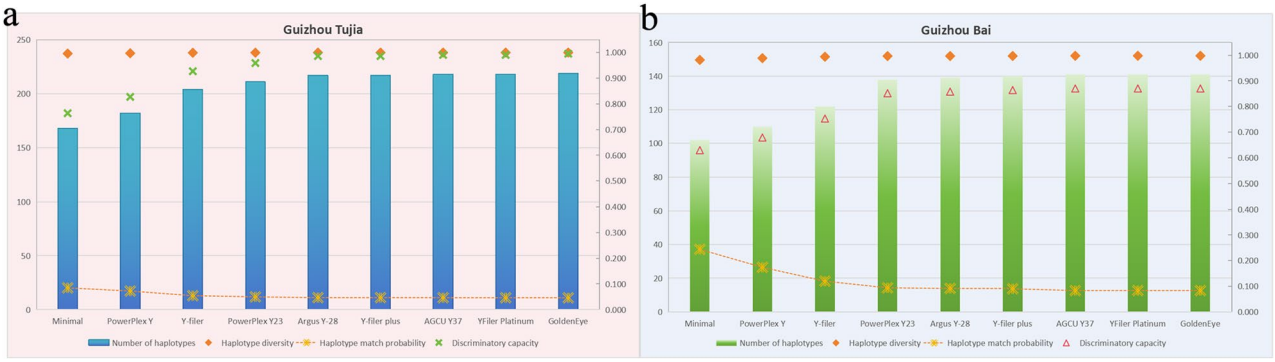

Fig. 2 Forensic efficiency comparisons of different Y-STR sets in the Guizhou Tujia (**a**) and Bai (**b**) populations, including haplotype number, haplotype diversity (HD), haplotype match probability (HMP), and discrimination capacity (DC)

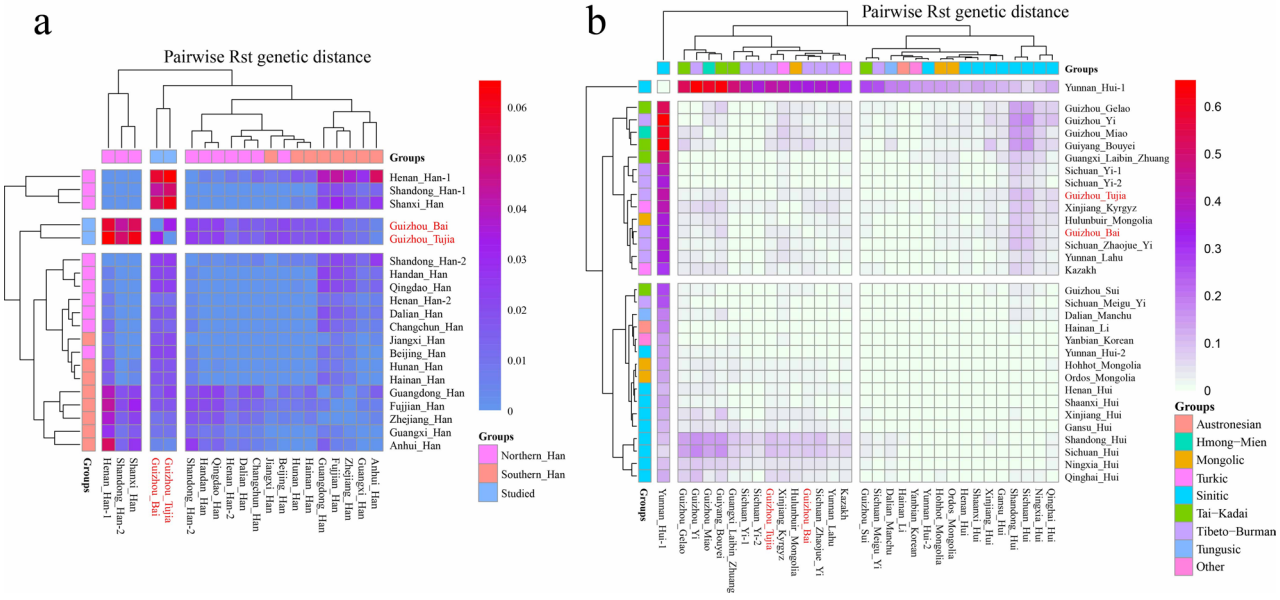

Fig. 3 Heatmap based on the paired Rst values among the Guizhou Tujia, Bai, and reference populations. **a** Northern and Southern Han populations. **b** East Asian minority populations

(Rst = 0.0034, 0.01005; Rst = 0.00467, 0.00911; respectively), while exhibiting a greater Rst to northern Han populations. The interpopulation Rst distance between the Tujia and Bai ethnic groups was comparatively large (Rst = 0.03301), exceeding their distances to several Southern Han references. The results indicated that the Tibeto-Burman-speaking populations in Guizhou are genetically closer to the Southern Han populations. Additionally, there was a significant genetic diversity difference among the ethnic minority groups in Guizhou, even within the same language family.

The Guizhou groups exhibited smaller Rst distances to geographically proximate minority populations from Sichuan, Guangxi, and Hainan, particularly with the Sichuan Yi population, a Tibeto-Burman-speaking population. The largest genetic distances were observed with Hui populations from Yunnan, Sichuan and Shandong. Noteworthy, the Rst between the Guizhou Tujia and Bai groups was higher than their distances to several minorities from other regions or language families. This may indicate genetic differentiation among populations within Guizhou.

Next, we conducted an MDS analysis based on pairwise Rst genetic distances, using datasets from Han and minority populations separately, as visualized in Fig. 4. Among Han references, Northern and Southern Han formed two distinct clusters with a clear separation (Fig. 4a). We observed that the Guizhou Tujia and Bai populations plotted in the right quadrants with Southern Han populations, but the two studied groups were mutually distant, consistent with previous research reports [45]. Specifically, the Guizhou Tujia was localized to the upper-right corner, positioned at some remove from the Southern Han cluster, whereas the Guizhou Bai clustered

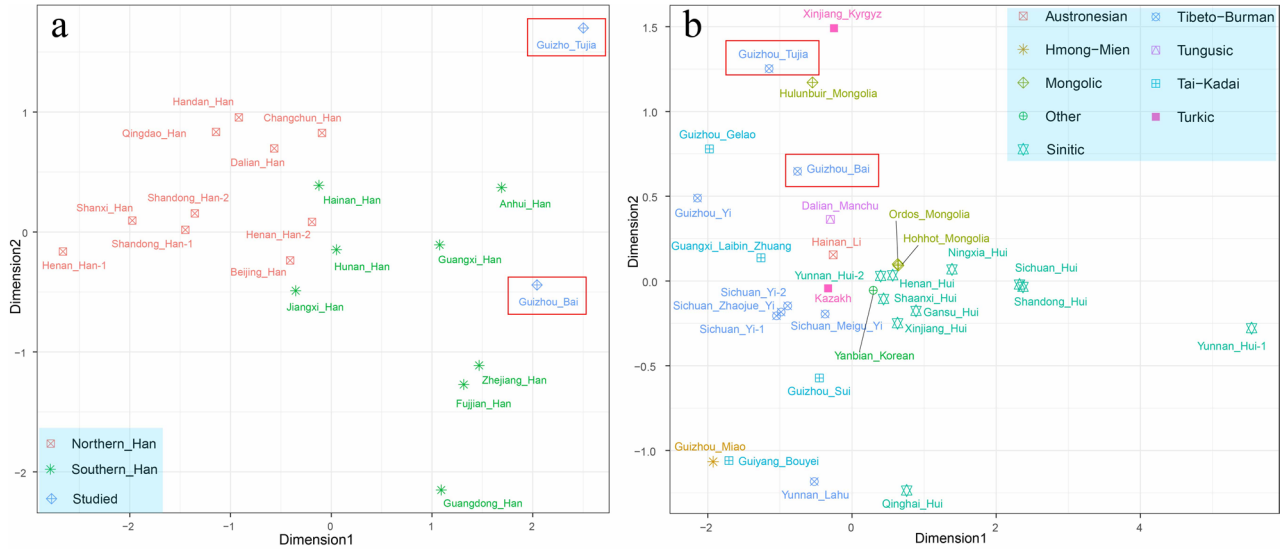

Fig. 4 Multidimensional scaling (MDS) plots showing genetic relationships among the Guizhou Tujia, Bai, and reference populations based on Rst genetic distances. **a** Han Chinese populations. **b** East Asian minority populations belonging to different linguistic families

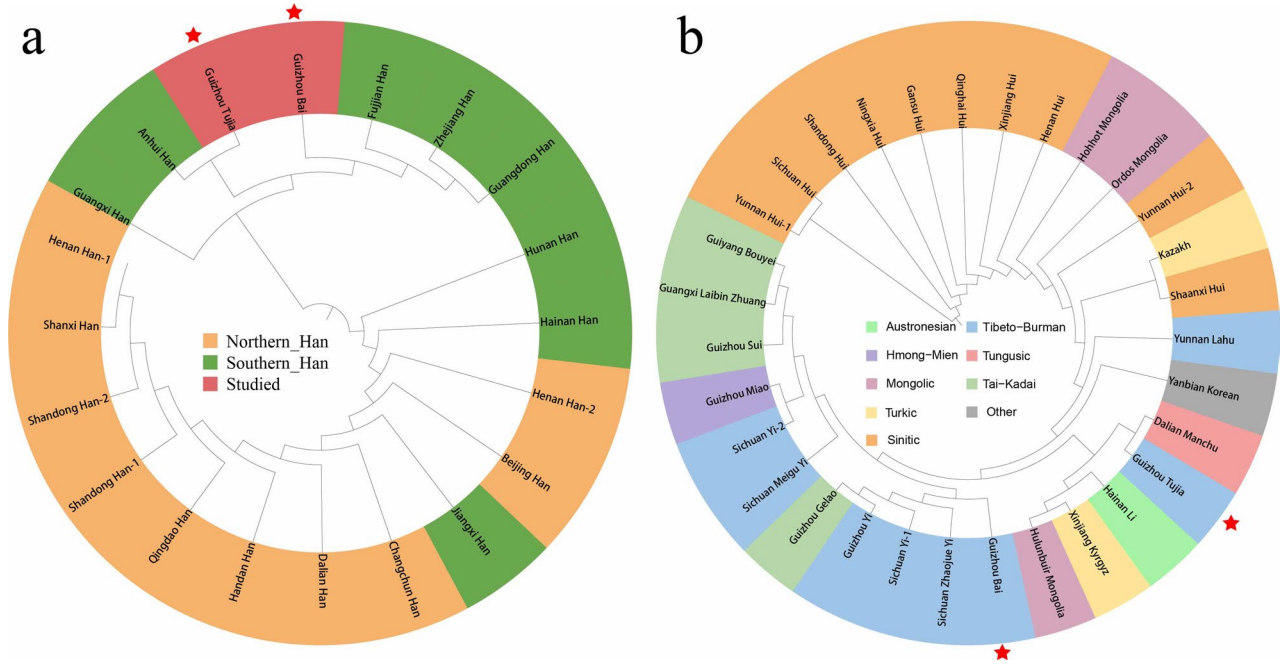

Fig. 5 Genetic similarity between the Guizhou Tujia and Bai populations and reference populations based on Rst genetic distances estimated from 27 Y-STRs. **a** Neighbor-joining tree of the Guizhou Tujia and Bai populations with Han Chinese populations; **b** Neighbor-joining tree of the studied populations with other reference minority populations

more closely with Southern Han populations. Figure 4b presents the MDS for the Guizhou Tujia and Bai alongside other ethnic minority groups. The Tujia and Bai were located in the upper-left quadrant together with the Guizhou Gelao and Yi, as well as the Kyrgyz (Xinjiang) and Mongolian (Hulunbuir), while the Bai population was closer to the Dalian Manchu, Guangxi Laibin Zhuang and Hainan Li groups. Most Sinitic-speaking Hui populations clustered together in the lower-right quadrant, whereas Tibeto-Burman-speaking Yi populations concentrated in the lower-left quadrant. No additional coherent clustering patterns were apparent among the remaining reference groups.

Finally, we constructed a neighbor-joining (NJ) tree based on pairwise Rst distances (Fig. 5). Consistent with the MDS analysis results, Northern and Southern Han formed distinct clades (Fig. 5a). The Guizhou Tujia and Bai formed a clade that, in turn, grouped with

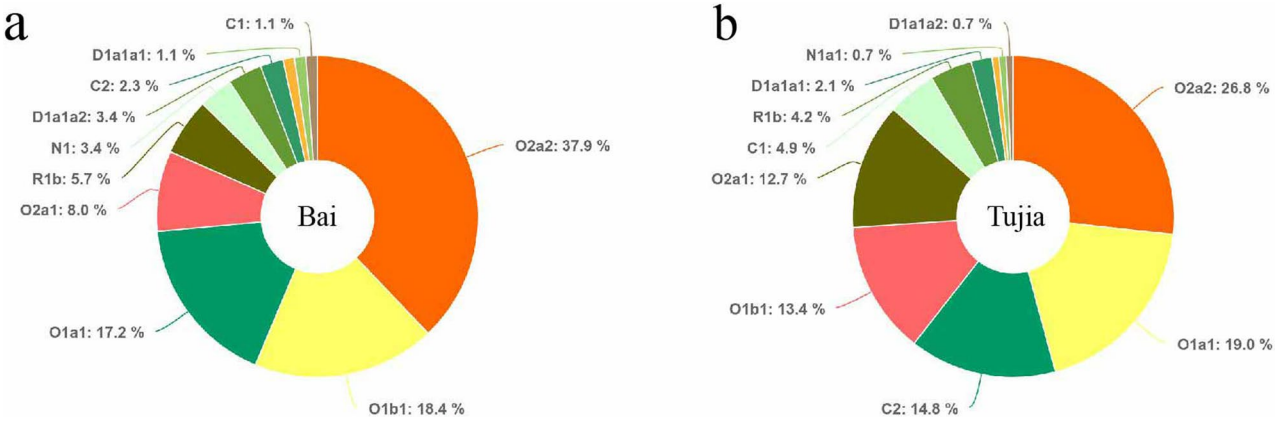

Fig. 6 Predicted Y-chromosomal haplogroup composition of the Guizhou Bai (**a**) and Tujia (**b**) populations inferred using the NevGen predictor

the Southern Han groups. In the minority-focused tree (Fig. 5b), the Guizhou Bai grouped with Tibeto-Burman-speaking populations, including the Sichuan Zhaojue Yi, Sichuan Yi-1 and Guizhou Yi populations. It appears that populations within similar language families exhibit a higher genetic affinity. Hui populations and Tai-Kadai-speaking groups likewise clustered within their respective language families. In contrast, the Guizhou Tujia formed an initial branch with the Dalian Manchu group, followed by the Hainan Li, Xinjiang Kyrgyz and Hulunbuir Mongolian populations.

Haplogroup prediction

Y-chromosomal haplogroups were predicted using the NevGen Y-DNA Haplogroup Predictor based on 30 Y-STR loci from the Goldeneye panel. Using a probability threshold of 0.80 for reliable assignment, 87 haplotypes from the Guizhou Bai group and 142 haplotypes from the Guizhou Tujia group reached or exceeded this cut-off and were therefore included in the haplogroup-based analyses. In the Bai group, the most frequent haplogroup was O2a2 (37.9%), followed by O1b1 (18.4%), O1a1 (17.2%), O2a1 (8.0%) and R1b (5.7%). Lower frequency haplogroups included N1 (3.4%), D1a1a2 (3.4%), C2 (2.3%), D1a1a1 (1.1%), and C1 (1.1%) (Fig. 6a). In the Tujia population, O2a2 (26.8%) was likewise predominant, followed by O1a1 (19.0%), C2 (14.8%), O1b1 (13.4%), and O2a1 (12.7%) (Fig. 6b). Several additional haplogroups, including C1, R1b, D1a1a1, N1a1 and D1a1a2 lineages, were observed at lower proportions. Overall, both groups are dominated by East Asian O lineages, whereas differences in the frequencies of C2 and minor N and R1b components already hint at subtle distinctions in their paternal genetic backgrounds. Additionally, we investigated the haplogroup affiliations of the genotypes missing DYS522 alleles. This analysis revealed no lineage- or haplogroup-specific patterns. Instead, the distribution of these missing profiles was consistent with the overall haplogroup composition of the populations, being concentrated in the high-frequency lineages such as O2a2, O1b1, and O1a1.

Discussion

This study demonstrates that the Goldeneye DNA Identification System Y-Plus kit provides overall robust forensic performance in the Guizhou Tujia and Bai populations, showing notably high genetic polymorphisms (GD > 0.5) in most markers. Particularly, DYS385, DYF387S1, DYS527, and DYF404S1 displayed the highest GD values among the markers, whereas the DYS645 locus exhibited limited genetic diversity (GD < 0.2) in both populations. Consistent with previous reports [7, 43, 45], DYF387S1 exhibited a triallelic pattern. Given that DYF387S1 is a multi-copy Y-STR, this is most plausibly explained by inter-copy variation and copy-number polymorphism rather than the emergence of a novel locus [16]. During kit evaluation, all loci generally showed high typing success except DYS522, which produced genotypes in about 63–69% of samples. To assess its influence, we compared DC with and without this locus and found only a slight (< 1%) increase when DYS522 was included. Although the occasional dropout of DYS522 has little effect on overall forensic efficiency, this instability should be taken into account in routine forensic evaluations.

In the Guizhou Tujia, increasing the number of Y-STR markers was associated with higher values of DC and HD, and lower HMP, compared with smaller marker sets. In contrast, the Guizhou Bai population showed little variation in these metrics among the Goldeneye Y Plus, Yfiler Plus, and AGCU Y37 panels, suggesting that the additional markers in the expanded panels provided limited incremental resolution for resolving the shared lineages specific to this group. Because HD provides a robust estimate of genetic variability that accounts for sample-size differences, this pattern most likely reflects population-specific haplotype composition rather than a

lack of statistical power. The Bai population exhibited a lower DC than the Tujia, which may result from several factors, including (i) higher haplotype redundancy within the Bai dataset, possibly due to cryptic relatedness or founder effects, (ii) differences in diversity at multi-copy or highly polymorphic loci, and (iii) sampling variance related to the smaller Bai sample size. Considering China's multiethnic complexity, future evaluations of Y-STR forensic parameters should encompass a broader set of Chinese populations and employ the largest feasible sample sizes to better capture rare variants and substantiate these regional differences.

Our study investigated the paternal genetic relationships of the Guizhou Tujia and Bai with Chinese Han populations and multiple minority groups across different language families. We observed a clear clustering divide between Southern and Northern Han populations, a pattern which is widely corroborated by previous research [46–48]. This substructure is largely attributable to Neolithic agricultural expansions from the Yellow and Yangtze River basins, followed by subsequent cultural admixture that together shaped Han genetic structure [49–51]. Consistent with this framework, prior studies have identified three major paternal subgroups within the Han (Northern, Southern, and Lingnan), further supporting the reliability of our findings [52]. As the demographically dominant population in China, the Han have substantially influenced the Y-chromosomal landscape of neighboring groups, with region-specific interactions between local Han and adjacent minorities [50]. Reflecting this geographic and historical context, the Guizhou Tujia and Bai in our dataset show stronger paternal affinity to Southern Han populations, such as those from Anhui, Guangxi and Zhejiang.

The minority panel MDS reveals two clustering patterns commonly observed in East Asian population genetics: (i) same ethnicity clustering (Yi and Hui groups) and (ii) spatially proximate clustering, whereby the Guizhou Tujia and Bai cluster with populations from neighboring provinces such as Sichuan, Guangxi, and Hainan. First, same ethnicity clustering reflects shared paternal lineages maintained by endogamy, clan-based social organization, and long-term historical continuity [53, 54]. Second, regional clustering reflects isolation by distance shaped by topography, where mountain ranges and major river systems restrict long-distance movement yet promote stepwise dispersal along valleys and basins, generating gradients and local clusters across adjacent prefectures and provinces [55]. These mechanisms are complementary rather than mutually exclusive. Methodologically, our inference is confined to the Y chromosome using a 27-locus Y-STR panel; thus, the structure recovered here should be regarded as indicative rather than definitive. More resolved reconstructions of population history, particularly disentangling recent male-mediated gene flow from deeper shared ancestry, will require genome-wide data (e.g., high-coverage autosomal SNPs or whole genome sequencing), integration of Y-SNP defined haplogroups to anchor lineages, and, where available, ancient DNA to provide temporal context.

Significant paternal divergence was observed between the Guizhou Tujia and Bai populations in our Y-STR dataset, even though both are classified as Tibeto-Burman-speaking groups. Historically, the Tujia population has received gene flow from northern and eastern sources, reflecting demographic inputs associated with past migrations from Central China and the upper Yangtze region [50, 56, 57]. In contrast, the Bai from western Guizhou exhibit closer affinities to Yunnan populations and display Y-chromosomal components linked to ancient southern East Asian lineages. These asymmetric gene-flow trajectories and deep-time population structures have likely shaped the distinct paternal signatures observed between the two groups. Because Y-STRs are particularly sensitive to recent demographic events, the differentiation may also capture localized founder effects and genetic drift within patrilineal clans [58, 59]. Collectively, ecological isolation, cultural endogamy, historical admixture, and lineage-specific drift provide an explanation for the observed paternal divergence between the Guizhou Tujia and Bai. This diversity is also reflected in the rich linguistic and cultural variety of the minority ethnic groups in Guizhou, underscoring the importance of future analyses integrating high-resolution Y-SNP data and genome-wide autosomal variation.

Furthermore, the paternal genetic landscapes of both the Guizhou Bai and Tujia populations are characterized by the predominance of haplogroup O2a2 and its subclades, reflecting a strong genetic signature of East Asian origin [60, 61]. O2 lineages are widely regarded as markers of the expansion of populations associated with Sino-Tibetan languages across East Asia. The notable presence of O1b1 and O1a1 in both groups further underscores their close genetic affinity with southern East Asian and Southeast Asian populations, highlighting historical connections and potential gene flow among regional ethnic groups. In addition to these predominant O lineages, both populations also carry minor components of C, D, N and R lineages, indicating more complex paternal contributions. In particular, the occurrence of R1b, although at low frequencies, may point to episodes of historical admixture with western Eurasian male lineages. Such signals could be related to ancient migration waves, localized gene flow, or long-distance contacts facilitated by trade networks traversing Southwest China [62]. The presence of such a spectrum of low-frequency haplogroups suggests that Guizhou has been influenced

by multiple prehistoric and historical migrations, rather than representing a completely isolated region [62].

Conclusions

The study reported the genetic polymorphisms of 43 Y-marker loci in Tujia and Bai male individuals from Guizhou Province in Southwest China. The Goldeneye Y-Plus kit demonstrated robust forensic efficiency with high haplotype diversity and discrimination capacity, confirming its suitability for forensic applications in these populations.

Population comparisons revealed that the Guizhou Tujia and Bai groups are genetically closer to Southern Han populations and other geographically neighboring ethnic groups. This pattern reflects shared ancestry and long-term regional gene flow within Southwest China. Furthermore, Y-STR haplotype predictions indicate that their paternal gene pools have been primarily shaped by indigenous East Asian lineages, with minor contributions from West Eurasian and other regional lineages.

Collectively, this study enriches the Y-STR database of Chinese Tibeto-Burman-speaking populations and refines our understanding of the paternal genetic structure, demographic history, and interpopulation dynamics of ethnic minorities in Guizhou.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-026-12719-6.

Supplementary Material 1.

Acknowledgements
The authors thank all the individuals who voluntarily participated in this study.

Authors' contributions
Conceptualization and resources, W.W. and M.Y.; methodology, W.W., J.Z. and N.A.; validation, P.Q., J.Z. and H.Z.; formal analysis, investigation, W.W., X.Z. and J.D.; visualization, writing—original draft preparation, W.W., H.Z. and X.Z.; writing—review and editing, W.W. and M.Y.; All authors have read and agreed to the published version of the manuscript.

Funding
Guizhou Medical University Doctor Foundation (QianJ-2022-PHD-078); Guizhou Scientific Basic Research Project, Qian Science Basic [2024] Youth 227; Guizhou College Students Innovation and Entrepreneurship Training Program (No. S202310660017).

Data availability
The data are provided within the manuscript or supplementary information files and have been deposited in the OMIX, China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences (https://ngdc.cncb.ac.cn/omix: accession no. OMIX015288).

Declarations

Ethics approval and consent to participate
Our study adhered to the Declaration of Helsinki. The study was approved by the ethics committee of Guizhou Medical University. All participants involved in the study provided written informed consent.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]School of Public Health and Management, Jiangsu Medical College, Yancheng, Jiangsu 224000, China
[2]College of Forensic Science, School of Medicine, Xi'an Jiaotong University, Xi'an, Shaanxi, China
[3]Department of Forensic Medicine, Guizhou Medical University, Guiyang, Guizhou 550004, China

Received: 14 August 2025 / Accepted: 27 February 2026
Published online: 25 March 2026

References

1. Kayser M. Forensic use of Y-chromosome DNA: a general overview. Hum Genet. 2017;136:621–35. https://doi.org/10.1007/s00439-017-1776-9.
2. Costa R, Fadoni J, Amorim A, Cainé L. Y-STR databases-application in sexual crimes. Genes. 2025;16:484. https://doi.org/10.3390/genes16050484.
3. SemanA, Sapawi AM, Salleh MZ. Towards Development of Clustering Applications for Large-Scale Comparative Genotyping and Kinship Analysis Using Y-Short Tandem Repeats. OMICS. 2015;19:361–7. . https://doi.org/10.1089/omi.2014.0136
4. Luis JR, Palencia-Madrid L, Deshpande K, Alfonso-Sanchez MA, Peña JA, de Pancorbo MM, et al. On the Y chromosome of chennai, tamil nadu and the indian subcontinent. Gene. 2023;859:147175. https://doi.org/10.1016/j.gene.2023.147175.
5. Ashirbekov Y, Seidualy M, Abaildayev A, Maxutova A, Zhunussova A, Akilzhanova A, et al. Genetic polymorphism of Y-chromosome in kazakh populations from southern Kazakhstan. BMC Genomics. 2023;24:649. https://doi.org/10.1186/s12864-023-09753-z.
6. Willuweit S, Roewer L. The new Y Chromosome Haplotype Reference Database. Forensic Sci Int Genet. 2015;15:43–8. https://doi.org/10.1016/j.fsigen.2014.11.024.
7. Thompson JM, Ewing MM, Frank WE, Pogemiller JJ, Nolde CA, Koehler DJ, et al. Developmental validation of the PowerPlex® Y23 System: a single multiplex Y-STR analysis system for casework and database samples. Forensic Sci Int Genet. 2013;7:240–50. https://doi.org/10.1016/j.fsigen.2012.10.013.
8. Zhu B, Zhang Y, Liu W, Meng H, Yuan G, Lv Z, et al. Genetic diversity and haplotype structure of 24 Y -chromosomal STR in C hinese H ui ethnic group and its genetic relationships with other populations. Electrophoresis. 2014;35:1993–2000. https://doi.org/10.1002/elps.201300574.
9. He G, Chen P, Zou X, Chen X, Song F, Yan J, et al. Genetic polymorphism investigation of the Chinese Yi minority using PowerPlex® Y23 STR amplification system. Int J Legal Med. 2017;131:663–6. https://doi.org/10.1007/s00414-017-1537-2.
10. Chen P, Han Y, He G, Luo H, Gao T, Song F, et al. Genetic diversity and phylogenetic study of the Chinese Gelao ethnic minority via 23 Y-STR loci. Int J Legal Med. 2018;132:1093–6. https://doi.org/10.1007/s00414-017-1743-y.
11. Yao H-B, Tang S, Yao X, Yeh H-Y, Zhang W, Xie Z, et al. The genetic admixture in Tibetan-Yi Corridor. Am J Phys Anthropol. 2017;164:522–32. https://doi.org/10.1002/ajpa.23291.
12. 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 Sci Int Genet. 2018;34:141–7. https://doi.org/10.1016/j.fsigen.2018.02.009.
13. Wang M, Wang Z, He G, Jia Z, Liu J, Hou Y. Genetic characteristics and phylogenetic analysis of three Chinese ethnic groups using the Huaxia Platinum System. Sci Rep. 2018;8:2429. https://doi.org/10.1038/s41598-018-20871-7.
14. Liu J, Wang Z, He G, Wang M, Hou Y. Genetic polymorphism and phylogenetic differentiation of the Huaxia Platinum System in three Chinese minority ethnicities. Sci Rep. 2019;9:3371. https://doi.org/10.1038/s41598-019-39794-y.
15. Guo J, Xu B, Li L, He G, Zhang H, Cheng H-Z, et al. Paternal Y chromosomal genotyping reveals multiple large-scale admixtures in the formation of Lolo-Burmese-speaking populations in southwest China. Ann Hum Biol. 2019;46:581–8. https://doi.org/10.1080/03014460.2019.1698655.

16. Chen J, Zhang H, Yang M, Wang R, Zhang H, Ren Z, et al. Genomic formation of Tibeto-Burman speaking populations in Guizhou, Southwest China. BMC Genomics. 2023;24:672. https://doi.org/10.1186/s12864-023-09767-7.
17. Zhang H, Huang X, Jin X, Ren Z, Wang Q, Yang M, et al. Comprehensive analyses of genetic diversities and population structure of the Guizhou Dong group based on 44 Y-markers. PeerJ. 2023;11:e16183. https://doi.org/10.7717/peerj.16183.
18. Nei M. Molecular evolutionary genetics. New York: Columbia University; 1987.
19. Nei M. Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci. 1973;70:3321–3. https://doi.org/10.1073/pnas.70.12.3321.
20. Gené M, Borrego N, Xifró A, Piqué E, Moreno P, Huguet E. Haplotype frequencies of eight Y-chromosome STR loci in barcelona (north-east Spain). Int J Legal Med. 1999;112:403–5. https://doi.org/10.1007/s004140050025.
21. Sun H, Su K, Fan C, Long F, Liu Y, Sun J, et al. Y-STRs' genetic profiling of 1953 individuals from two Chinese Han populations (Guizhou and Shanxi). Forensic Sci Int Genet. 2019;38:e8–10. https://doi.org/10.1016/j.fsigen.2018.10.011.
22. Chen P, Wu J, Luo L, Gao H, Wang M, Zou X, et al. Population Genetic Analysis of Modern and Ancient DNA Variations Yields New Insights Into the Formation, Genetic Structure, and Phylogenetic Relationship of Northern Han Chinese. Front Genet. 2019;10:1045. https://doi.org/10.3389/fgene.2019.01045.
23. Wang C-Z, Wei L-H, Zhang J-S, Yu X-E, Bai R-F, Li H, et al. Forensic characteristics and genetic substructure analysis of the Handan Han population, Northern China. Ann Hum Biol. 2023;50:123–5. https://doi.org/10.1080/03014460.2023.2181985.
24. Fan G-Y, Jiang D-Z, Jiang Y-H, Song W, He Y-Y, Wuo NA. Phylogenetic analyses of 41 Y-STRs and machine learning-based haplogroup prediction in the Qingdao Han population from Shandong province, Eastern China. Ann Hum Biol. 2023;50:35–41. https://doi.org/10.1080/03014460.2023.2168057.
25. Liu Y, Jin H-Y, Song D-L, Wang C-Z, Shi M-S, Wang C-C. 27 Y-chromosomal STR haplotypic structure for the Chinese Han population from Changchun, Northeastern China. Leg Med (Tokyo). 2021;48:101813. https://doi.org/10.1016/j.legalmed.2020.101813.
26. Chen W-Q, Feng Z, Jin W, Zhang Y-J, Wang Q, Li P, et al. Genetic analysis of 17 Y-STR loci from 1026 individuals of Han populations in Jilin Province, Northeast China. Int J Legal Med. 2018;132:1309–11. https://doi.org/10.1007/s00414-018-1778-8.
27. Li L, Yu G, Li S, Jin L, Yan S. Genetic analysis of 17 Y-STR loci from 1019 individuals of six Han populations in East China. Forensic Sci Int Genet. 2016;20:101–2. https://doi.org/10.1016/j.fsigen.2015.10.007.
28. Zhao M, Cui W, Zhang Y, Lan Q, Zhu B. Haplotypic polymorphisms and forensic applications in Chinese Hunan Han population based on a series of Y-STR loci: a perspective of paternal inheritance. Ann Hum Biol. 2023;50:52–5. https://doi.org/10.1080/03014460.2023.2171120.
29. Fan H, Zeng Y, Wu W, Liu H, Xu Q, Du W, et al. The Y-STR landscape of coastal southeastern Han: Forensic characteristics, haplotype analyses, mutation rates, and population genetics. Electrophoresis. 2021;42:1578–93. https://doi.org/10.1002/elps.202100037.
30. Lang M, Liu H, Song F, Qiao X, Ye Y, Ren H, et al. Forensic characteristics and genetic analysis of both 27 Y-STRs and 143 Y-SNPs in Eastern Han Chinese population. Forensic Sci Int Genet. 2019;42:e13–20. https://doi.org/10.1016/j.fsigen.2019.07.011.
31. Lin H, Ye Q, Tang P, Mo T, Yu X, Tang J. Analyzing genetic polymorphism and mutation of 44 Y-STRs in a Chinese Han population of Southern China. Leg Med (Tokyo). 2020;42:101643. https://doi.org/10.1016/j.legalmed.2019.101643.
32. Wang C-Z, Zhang J-S, Li X-B, Bai R-F, Shi M-S, Wang C-C. Haplotype analysis of 36 Y-STR loci in a Chinese Han population from Anhui Province, Eastern China. Int J Legal Med. 2020;134:2063–5. https://doi.org/10.1007/s00414-020-02321-0.
33. Xie M, Song F, Li J, Lang M, Luo H, Wang Z, et al. Genetic substructure and forensic characteristics of Chinese Hui populations using 157 Y-SNPs and 27 Y-STRs. Forensic Sci International: Genet. 2019;41:11–8. https://doi.org/10.1016/j.fsigen.2019.03.022.
34. Fan G-Y, An Y-R, Peng C-X, Deng J-L, Pan L-P, Ye Y. Forensic and phylogenetic analyses among three Yi populations in Southwest China with 27 Y chromosomal STR loci. Int J Legal Med. 2019;133:795–7. https://doi.org/10.1007/s00414-018-1984-4.
35. Liu Y, Wang C, Zhou W, Li X-B, Shi M, Bai R, et al. Haplotypes of 27 Y-STRs analyzed in Gelao and Miao ethnic minorities from Guizhou Province, Southwest China. Forensic Sci Int Genet. 2019;40:e264–7. https://doi.org/10.1016/j.fsigen.2019.03.002.
36. Luo Y, Wu Y, Qian E, Wang Q, Wang Q, Zhang H, et al. Population genetic analysis of 36 Y-chromosomal STRs yields comprehensive insights into the forensic features and phylogenetic relationship of Chinese Tai-Kadai-speaking Bouyei. PLoS ONE. 2019;14:e0224601. https://doi.org/10.1371/journal.pone.0224601.
37. Ji J, Ren Z, Zhang H, Wang Q, Wang J, Kong Z, et al. Genetic profile of 23 Y chromosomal STR loci in Guizhou Shui population, southwest China. Forensic Sci Int Genet. 2017;28:e16–7. https://doi.org/10.1016/j.fsigen.2017.01.010.
38. Fan H, Wang X, Chen H, Zhang X, Huang P, Long R, et al. Population analysis of 27 Y-chromosomal STRs in the Li ethnic minority from Hainan province, southernmost China. Forensic Sci Int Genet. 2018;34:e20–2. https://doi.org/10.1016/j.fsigen.2018.01.007.
39. Han Y, Li L, Liu X, Chen W, Yang S, Wei L, et al. Genetic analysis of 17 Y-STR loci in Han and Korean populations from Jilin Province, Northeast China. Forensic Sci International: Genet. 2016;22:8–10. https://doi.org/10.1016/j.fsigen.2016.01.003.
40. Song F, Song M, Luo H, Xie M, Wang X, Dai H, et al. Paternal genetic structure of Kyrgyz ethnic group in China revealed by high-resolution Y-chromosome STRs and SNPs. Electrophoresis. 2021;42:1892–9. https://doi.org/10.1002/elps.202100142.
41. Luo H, Song F, Zhang L, Hou Y. Genetic polymorphism of 23 Y-STR loci in the Zhuang minority population in Guangxi of China. Int J Legal Med. 2015;129:737–8. https://doi.org/10.1007/s00414-015-1175-5.
42. Gao T, Yun L, Gao S, Gu Y, He W, Luo H, et al. Population genetics of 23 Y-STR loci in the Mongolian minority population in Inner Mongolia of China. Int J Legal Med. 2016;130:1509–11. https://doi.org/10.1007/s00414-016-1433-1.
43. KumarS,Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 2016;33:1870–4. . https://doi.org/10.1093/molbev/msw054
44. Zhou Y, Shao C, Li L, Zhang Y, Liu B, Yang Q, et al. Genetic analysis of 29 Y-STR loci in the Chinese Han population from Shanghai. Forensic Sci Int Genet. 2018;32:e1–4. https://doi.org/10.1016/j.fsigen.2017.11.003.
45. Luo Y, Chi A, Guo X, Song J, Zhang X, Dong Y, et al. Genetic mixed diversity landscape in the paternal lineages of 11 populations inhabiting southwest China according to the analysis of 25 Y-STRs. BMC Ecol Evol. 2025;25:61. https://doi.org/10.1186/s12862-025-02400-z.
46. Chiang CWK, Mangul S, Robles C, Sankararaman S. A comprehensive map of genetic variation in the world's largest ethnic group—han chinese. Mol Biol Evol. 2018;35:2736–50. https://doi.org/10.1093/molbev/msy170.
47. Chen J, Zheng H, Bei J-X, Sun L, Jia W, Li T, et al. Genetic structure of the han chinese population revealed by genome-wide SNP variation. Am J Hum Genet. 2009;85:775–85. https://doi.org/10.1016/j.ajhg.2009.10.016.
48. Zhu X, Jin X, Liu J, Yang L, Zou L-X, Li C-X, et al. Paternal genetic structure analysis of the modern han populations based on Y-SNP and Y-STR. Yi Chuan = Hered. 2024;46:149–67. https://doi.org/10.16288/j.yczz.23-260.
49. Du P, Zhu K, Wang M, Sun Z, Tan J, Sun B, et al. Genomic dynamics of the lower yellow river valley since the early neolithic. Curr Biol. 2024;34:3996–e400611. https://doi.org/10.1016/j.cub.2024.07.063.
50. Wang C-C, Yeh H-Y, Popov AN, Zhang H-Q, Matsumura H, Sirak K, et al. Genomic insights into the formation of human populations in east Asia. Nature. 2021;591:413–9. https://doi.org/10.1038/s41586-021-03336-2.
51. Ji Z, Chen K, Zheng J, Qin C, Cui S, Shen Q, et al. Genomic formation of lower yellow river populations in the han dynasty. BMC Biol. 2025;23:260. https://doi.org/10.1186/s12915-025-02377-7.
52. Wang Z, Liu K, Yuan H, Duan S, Liu Y, Luo L, et al. YanHuang paternal genomic resource suggested a weakly-differentiated multi-source admixture model for the formation of han's founding ancestral lineages. Genomics Proteom Bioinf. 2025; [Epub ahead of print] https://doi.org/10.1093/gpbjnl/qzaf049.
53. López S, Thomas MG, Van Dorp L, Ansari-Pour N, Stewart S, Jones AL, et al. The genetic legacy of zoroastrianism in Iran and India: insights into population structure, gene flow, and selection. Am J Hum Genet. 2017;101:353–68. https://doi.org/10.1016/j.ajhg.2017.07.013.
54. Marchani EE, Watkins WS, Bulayeva K, Harpending HC, Jorde LB. Culture creates genetic structure in the caucasus: autosomal, mitochondrial, and Y-chromosomal variation in daghestan. BMC Genet. 2008;9:47. https://doi.org/10.1186/1471-2156-9-47.
55. Li Y-C, Ye W-J, Jiang C-G, Zeng Z, Tian J-Y, Yang L-Q, et al. River valleys shaped the maternal genetic landscape of han chinese. Mol Biol Evol. 2019;36:1643–52. https://doi.org/10.1093/molbev/msz072.
56. Chen J, Wang M, Duan S, Yang Q, Liu Y, Zhao M, et al. Genetic history and biological adaptive landscape of the tujia people inferred from shared

haplotypes and alleles. Hum Genomics. 2024;18:104. https://doi.org/10.1186/s40246-024-00672-y.
57. Wan W, Huang H, Hu C, Zhang X, Zhou Q, Yang M, et al. Ancient genomes in southwest China revealed genetic interactions among diverse populations in the historical period. BMC Biol. 2025;23:174. https://doi.org/10.1186/s12915-025-02299-4.
58. Di Giacomo F, Luca F, Anagnou N, Ciavarella G, Corbo RM, Cresta M, et al. Clinal patterns of human Y chromosomal diversity in continental italy and greece are dominated by drift and founder effects. Mol Phylogenet Evol. 2003;28:387–95. https://doi.org/10.1016/S1055-7903(03)00016-2.
59. Chiaroni J, Underhill PA, Cavalli-Sforza LL. Y chromosome diversity, human expansion, drift, and cultural evolution. Proc Natl Acad Sci. 2009;106:20174–9. https://doi.org/10.1073/pnas.0910803106.
60. Deng W, Shi B, He X, Zhang Z, Xu J, Li B, et al. Evolution and migration history of the chinese population inferred from chinese Y-chromosome evidence. J Hum Genet. 2004;49:339–48. https://doi.org/10.1007/s10038-004-0154-3.
61. Yan S, Wang C-C, Li H, Li S-L, Jin L, Genographic Consortium. An updated tree of Y-chromosome haplogroup O and revised phylogenetic positions of mutations P164 and PK4. Eur J Hum Genet: EJHG. 2011;19:1013–5. https://doi.org/10.1038/ejhg.2011.64.
62. Zhu K, Hu C, Yang M, Zhang X, Guo J, Xie M, et al. The demic diffusion of han culture into the yunnan-guizhou plateau inferred from ancient genomes. Natl Sci Rev. 2024;11:nwae387. https://doi.org/10.1093/nsr/nwae387.

Publisher's note