Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Analysis of genetic diversity and population Structure of Gleditsia sinensis based on SLAF-Seq

**Bing Yang**

136390604480163.com

**Jun Luo**

**Xiaoyong Dai**

**Research Article**

**Keywords:** Gleditsia sinensis, SLAF-seq, SNP, genetic diversity

**Posted Date:** April 1st, 2025

**DOI:** https://doi.org/10.21203/rs.3.rs-6259139/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

**Additional Declarations:** No competing interests reported.

**Version of Record:** A version of this preprint was published at Genetic Resources and Crop Evolution on December 9th, 2025. See the published version at https://doi.org/10.1007/s10722-025-02690-8.

# Abstract

*Gleditsia sinensis* Lam. has significant medicinal and economic value. By utilizing SLAF-Seq technology, we developed single nucleotide polymorphism (SNP) markers to analyze the genetic diversity of this species, providing a basis for the conservation and utilization of germplasm resources. In this study, we conducted SNP marker development on 159 *G. sinensis* samples from Guizhou Province, identifying a total of 132,709 population SNPs. The number of SNPs per sample ranged from 27,292 to 76,870, with heterozygosity rates ranging from 4.17–15.92%. The analysis of population genetic structure indicated that these 159 *G. sinensis* samples could be divided into five groups. The genetic distances among samples from the same provenance were close, while gene flow was observed among germplasm resources from different geographical populations. The nucleotide diversity level of population Q5 was relatively high, and the Tajima's D test suggested that rare alleles were scarce within the populations, with the majority of genes being neutral, indicating a state of equilibrium. This study provides fundamental data for the collection of *G. sinensis* germplasm resources and the breeding of superior varieties.


# Introduction

*Gleditsia sinensis* Lam. is a species that belongs to the genus *Gleditsia* within the subfamily Caesalpinioideae of the family Leguminosae. This species is valued for its medicinal and culinary applications.The primary utilizable components of the plant include thorns, pods, and seeds (commonly referred to as "soap bean kernels") Each of these components possesses considerable medicinal and economic value. (Xiao et al., 2023). The thorns are used in traditional medicine (Shu et al., 2023), the pods are rich in saponins and serve as natural detergents, while the seeds are edible. Additionally, the durable wood of *G. sinensis* is iparticularly well-suited for both domestic and artisanal applications.This species is widely distributed across 19 provinces in China. MaxEnt and ArcGIS analyses suggest that *G. sinensis* possesses a wide range of potentially suitable habitats across China, with medium- and high-suitability zones covering approximately 13.76% of the country's total area (Yang et al., 2022).

Investigating and collecting elite germplasm resources is essential for promoting the development of the *G. sinensis* industry and enhancing genetic improvement. The genetic diversity of *G. sinensis* populations is intricately linked to their geographic distribution as well as anthropogenic activities (Li et al., 2017).Research on the germplasm of Guizhou has demonstrated significant phenotypic variation both among and within populations (Tian et al., 2022).Pod morphological indices exhibit an increasing trend with elevation, whereas seed size shows a decreasing pattern. Additionally, seeds demonstrate a gradual enlargement trend from the south to the north and from the east to the west across Guizhou (Tian, 2022). Seedling height growth across ten populations from six southern provinces of China exhibits a significant negative correlation with both latitude and annual average temperature, indicating the presence of a latitudinal cline.(Li, 2013).SRAP (sequence-related amplified polymorphism) marker analysis of 203 samples from 14 natural populations revealed genetic distances ranging from 0.0504 to 0.3596, with significant positive correlations between geographic and genetic distances (Tian et al., 2023).Amplification techniques such as RASP (Fan et al., 2017), SRAP (Zhang et al., 2017), and SCoT (Zhang et al., 2017) have demonstrated a high level of polymorphism across 18 germplasm resources of *G. sinensis*.

Simplified genome sequencing provides a technical basis for elucidating patterns of species population distribution and facilitating genetic enhancement.SLAF-Seq (Specific-Locus Amplified Fragment Sequencing) identifies approximately one tag per 10 kb with a uniform distribution while avoiding repetitive sequences. This method is more cost-effective compared to Genotyping-by-Sequencing (GBS) or Restriction-site Associated DNA Sequencing (RAD-seq) (Chen et al., 2022; Sun et al., 2013). Due to its high efficiency in the development of SNP markers, SLAF-Seq has been extensively utilized across a variety of plant species.(Shen et al., 2024; Li et al., 2024). In this study, we utilized SLAF-Seq to generate SNP markers for 159 Gleditsia sinensis germplasms collected from Guizhou Province. This analysis aimed to assess

their genetic diversity and provides a foundational basis for the selection of germplasm resources and the identification of cultivars.

## Materials and Methods

# Plant Materials

A total of 159 leaf samples of *G. sinensis* were collected from eight distinct regions across Guizhou Province (Table 1). Healthy leaves were immediately flash-frozen in liquid nitrogen and subsequently stored at -80°C. DNA extraction was performed using the CTAB method, followed by quantification via spectrophotometry and verification through 1.5% agarose gel electrophoresis. Qualified DNA samples were preserved at -20°C for subsequent experiments.

Table 1
Leaf collection information of *G. sinensis* population in Guizhou Province

| Group | Region | Collection Sites | Longitude (°E) | Latitude (°N) | Altitude (m) | No. of Individuals |
|---|---|---|---|---|---|---|
| QN | Qiannan | Guiding,Huishui | 106.18 ~ 106.93 | 25.89 ~ 26.73 | 928.83 ~ 1390.4 | 13 |
| QDN | Qiandongnan | Leishan,Majiang,Taijiang | 106.88 ~ 108.07 | 26.40 ~ 28.73 | 613.45 ~ 1330.2 | 28 |
| AS | Anshun | Pingba | 106.22 ~ 106.89 | 26.35 ~ 26.47 | 1295.3 ~ 1398.0 | 12 |
| QXN | Qianxinan | Puan,Qinglong,Anlong,Ceheng,Xingren | 105.00 ~ 105.98 | 24.91 ~ 25.96 | 751.28 ~ 1582.2 | 20 |
| BJ | Bijie | Zhijin,Dafang | 105.38 ~ 106.03 | 26.49 ~ 27.31 | 1199.0 ~ 1703.0 | 22 |
| LPS | Liupanshui | Liupanshui | 104.62 ~ 105.63 | 26.17 ~ 26.81 | 1215.0 ~ 1935.7 | 19 |
| TR | Tongren | Sinan,Dejiang,Yinjiang,Shiqian | 107.97 ~ 108.61 | 27.48 ~ 28.30 | 414.7 ~ 1012.0 | 21 |
| ZY | Zunyi | Huichuan,Honghuagang,Meitan,Wuchuan | 106.67 ~ 107.90 | 27.56 ~ 28.80 | 712.0 ~ 1330.2 | 24 |

## Restriction Digestion and High-Throughput Sequencing

Genomic DNA was digested using RsaI and HaeIII, resulting in the selection of fragments ranging from 364 to 414 bp as SLAF tags. Libraries were constructed through A-tailing, adapter ligation, PCR amplification, and size selection (Kozich et al., 2013), followed by sequencing on an Illumina platform. The data were processed with FASTQ and BMK_Control for quality assessment.

## SNP Marker Statistics and Genetic Diversity Analysis

SNP markers were developed using the highest-depth sequence from each SLAF tag as a reference. Reads were aligned with BWA, and SNPs were identified using GATK (v3.8) and SAMtools (v1.9). Genome-wide SNPs were filtered based on minor allele frequency (MAF: 0.05) and site integrity (INT: 0.5) to retain high-confidence SNPs for subsequent analyses.

For phylogenetic analysis, a neighbor-joining tree was constructed using MEGA X (Tamura et al., 2011), employing the Kimura two-parameter model with 1,000 bootstrap replicates. Population structure was inferred through ADMIXTURE, testing subgroup numbers (K values) ranging from 1 to 10, followed by clustering and cross-validation.

Principal component analysis (PCA) was performed on the SNP data using EIGENSOFT software (version 6.0; Price et al., 2006). Genetic diversity parameters were assessed utilizing VCFtools.

## Results and Analysis

# SLAF Marker Development

The SLAF-Seq analysis of 159 *G. sinensis* samples generated a total of 805 Mb of sequencing data, achieving an average Q30 score of 93.50% and a GC content of 38.42%. A comprehensive set of 844,847 SLAF tags was obtained, with an average sequencing depth of 12.22x. Utilizing the Genome Analysis Toolkit (GATK), we identified 132,709 high-quality SNPs across the samples, with individual sample counts ranging from 27,292 to 76,870 SNPs and heterozygosity rates varying between 4.17% and 15.92%.

# Population Genetic Structure Analysis

Population structure analysis of 159 individuals, with K values preset from 1 to 10, revealed the lowest cross-validation error at K = 5 (Fig. 4a), indicating five distinct genetic groups. Group Q1 included CH-1, CH-2, CH-7, HC-34, HC-39, MT-8, MT-8-1, PA-6, TZ-1, and TZ-6; Q2 comprised DF-5, HC-5, HS-1, HS-15, HS-17, LPS-1, LPS-20, LPS-23, and LPS-69; Q3 encompassed DF-7, DJ-11, DJ-22, DJ-24, DJ-5, DJ-6, and HHG-15; Q4 included AL-2, DF-2, GD-12, GD-17, GD-3, HC-36, and HHG-13; Q5 consisted of DF-12, DF-9, DQ, GD-11, HC-9, HHG-03, HS-12, HS-4, HS-6, HS-9, LPS-12, LPS-21, LPS-35, LPS-4, LPS-56, LPS-61, LPS-62, LS-11, LS-14, MJ-1, MJ-21, MJ-27, MJ-32, MJ-56, MT-9, PA-3, PB-12, PB-3, PB-5, PB-9, QL-6, QL-9, SN-16, ZJ-14, ZJ-17, ZJ-2, ZJ-28, ZJ-38, and ZJ-6.

## Phylogenetic Tree Analysis

A neighbor-joining phylogenetic tree constructed using MEGA X (Fig. 2a) revealed close genetic distances among samples originating from the same geographic locations. Conversely, principal component analysis (PCA) of population structure (Fig. 2b) indicated potential gene flow between populations.

# Genetic Diversity Analysis

Individual heterozygosity ranged from − 0.41758 to 0.86718, with mean values for Q1–Q5 being 0.2047, 0.3696, 0.1933, 0.1600, and 0.1168, respectively. Nucleotide diversity (Pi) analysis (Fig. 3a) indicated a slightly higher diversity in Q5 (mean Pi = 0.0285), which included individuals from Zhijin (ZJ-14, ZJ-17, ZJ-2, ZJ-28, ZJ-38, ZJ-6), Liupanshui (LPS-12, LPS-21, LPS-35, LPS-4, LPS-56, LPS-61, LPS-62), and Majiang (MJ-1, MJ-21, MJ-27, MJ-32, MJ-56). Tajima's D values (Fig. 3b) averaged at 0.359 for Q5, surpassing those of the other groups examined. Pairwise Fst values Q1, Q2, Q3, Q4, and Q5 were 0.07, 0.026, 0.032, and 0.032, respectively, suggesting moderate genetic differentiation among these groups.

## Discussion

With the advancement of high-throughput sequencing technologies, SLAF-Seq, which utilizes second-generation sequencing methods, has demonstrated its efficacy in developing reliable SNP and InDel markers. This approach offers significant advantages for germplasm classification, studies on population evolution, and gene association analyses. For example, simplified genome sequencing has been employed to differentiate 12 clones of Camellia oleifera and to construct SNP fingerprint profiles (Liao et al., 2024). Similarly, SLAF-Seq facilitated the classification of 195 pepper germplasms into three distinct subgroups (Chen et al., 2023). In this study, SLAF-Seq was utilized for the first time to evaluate the genetic diversity of 159 Gleditsia sinensis samples collected from Guizhou Province. Using GATK software, a total of 132,709 high-quality SNPs were identified; within individual samples, SNP counts ranged from 27,292 to 76,870 with heterozygosity rates varying between 4.17% and 15.92%.

Population genetic structure analysis indicated that the 159 *G. sinensis* individuals could be categorized into five distinct groups. Genetic distances were found to be closer among individuals sharing the same geographic origin, suggesting potential gene flow between populations. The mean heterozygosity values for groups Q1 through Q5 were 0.2047, 0.3696, 0.1933, 0.1600, and 0.1168, respectively.Population Q5 exhibited a marginally elevated level of nucleotide diversity (Pi = 0.0285) in comparison to the other groups. Furthermore, Tajima's D, which quantifies intraspecific polymorphism and serves as an indicator for neutral evolution, was significantly higher in Q5 than in the other populations.All groups exhibited positive Tajima's D values, indicating an excess of intermediate-frequency alleles. This phenomenon may be attributed to bottleneck effects, population structure dynamics, or balancing selection (Li et al., 2020). Simplified genome sequencing of the 159 *G. sinensis* samples revealed moderate to strong genetic differentiation among them, with pairwise Fst values ranging from 0.07 to 0.80 (Yin et al., 2023). Notably, the average Fst values recorded at 0.07, 0.026, 0.032, and 0.032 between Q5 and the other groups (Q1–Q4), respectively. this indicates relatively low levels of genetic differentiation overall.

# Declarations


### Acknowledgements

We extend our gratitude to the researchers involved in the Project for their dedication and hard work. Additionally,We thank Biomarker Technologies Co.,Ltd for assisting in sequencing and/or bioinformatics analysis.

### Funding

This project was supported by The Guizhou Provincial Basic Research Program (No.ZK[2022]-242) and The Guizhou Forestry Seedling Cultivation Project (No. 2024-ZM-19).


### Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

### Author Contributions

Bing Yang and Jun Luo designed and executed the experiments, analyzed the data, and drafted the manuscript. Yayan Zhu contributed to both the experimental design and data analysis. Bing Yang and Xiaoyong Dai supervised the project, provided guidance on analyses, and revised the manuscript. All authors reviewed and approved the final version.

### Data Availability

The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive in National Genomics Data Center, China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences (GSA:CRA023840) that are publicly accessible at https://ngdc.cncb.ac.cn/gsa, (accessed on 18 March 2026).

# References


1. Chen Xiaocui,Zhang Xiaowei,Li Hui,Qiu Huarong,Tang Xiangqun,Luo Xirong,Yang Hong,Qin Cheng.SNP Sites Development by Specific Length Amplification Fragment Sequencing (SLAF-seq) and Genetic Analysis in Pepper.Molecular Plant Breeding.1-212
2. Chen Z, He Y, Iqbal Y, Shi Y, Huang H, Yi Z: Investigation of genetic relationships within three Miscanthus species using SNP markers identified with SLAF-seq. BMC genomics 2022, 23:1–11.
3. Fan Dingchen,Zhang Anshi,Liu Ying,Luo YangGenetic Diversity and Fingerprints of Gledіysia sinensis
4. Germplasms Based on RSAP.Journal of Henan Agricultural Sciences 2017, 46(11):103–107.
5. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD: Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Applied and environmental microbiology 2013, 79(17):5112–5120.
6. Li W, Liu L, Wang Y, Zhang Q, Fan G, Zhang S, Wang Y, Liao K: Genetic diversity, population structure, and relationships of apricot (Prunus) based on restriction site-associated DNA sequencing. Horticulture Research 2020, 7.
7. Li Wei,2013,Research on Genetic Resources Evaluation of Gleditsia sinensis in Sourth China,Doctoral dissertation,Chinese Academy Forestry,Mentor:Zheng Yongqi
8. LI Wei,LIN Fu-rong,ZHENG Yong-qi,SUN Sheng.Genetic Diversity of Ten Gleditsia sinensis Populations from Southern China.Forest Research 2017,30(01):46–52.
9. Li Xue-jun,Wang Bu-tian,Zhao Meng,Shi Xue-dong,Chen Zhi-hua,Xie Chun,Lu Yun-feng,Dong Xiang-shu,Ma Yin-bo,Ge Yu.Genetic diversity of Coffea arabica based on SLAF-Seq technology.Journal of Southern Agriculture 2024, 55(06):1583–1593.
10. Liao Hongze,Sun Manman,Huang Xiaojuan,Hao Bingqing,Sun Jiaxing,Jiang Zepeng,Wang Dongxue,Liu Kai.Development of SNP markers and construction of DNA fingerprint in teaoil Camellia germplasms based on simplified genome sequencing.Journal of Central South University of Forestry & Technology 2024, 44(04):128–137.
11. 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(8):904–909.
12. Shen Chaogui,Lai Ruilian,Chen Jin,Feng Xin,Chen Yiting,Wei Xiaoxia,Wu Rujian.SNP Markers Development and Genetic Relationship Identification of Germplasm Resources of Canarium album Based on SLAF-Seq Technology.Fujian Journal of Agricultural Sciences2024, 39(05):557–562.
13. Shu Chengjie,Li Zhuohang,Chen Diansong,Ma Ling,Zhou Xinyu,Zhao Jing,Wang Miao,Ma Shihong.Total Flavonoids, Anti-mite Activity and Antioxidant Activity ofthe Extract of Gleditsia Sinensis Thron.Chinese Wild Plant Resources 2022, 33(05):1201–1204.
14. Sun X, Liu D, Zhang X, Li W, Liu H, Hong W, Jiang C, Guan N, Ma C, Zeng H: SLAF-seq: an efficient method of large-scale de novo SNP discovery and genotyping using high-throughput sequencing. PloS one 2013, 8(3):e58700.
15. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular biology and evolution 2011, 28(10):2731–2739.
16. Tian Honghong,2022,Study on the genetic diversity of wild soap pod resources in Guizhou Province,Master's Thesis,Mentor:Zhao Yang
17. Tian Honghong,Yang Ju,Lu Chunyun,XiaoO Feng,Zhao Yang.Phenotypic Diversity Analysis of Provenances of Natural Gleditsia sinensis in Guizhou.Acta Botanica Boreali-occidentalis 2022, 42(11):1927–1935
18. Yang Ye,Liu Xiao-long,Lu Xiang,Sun Qing-wen.Ecological suitability regionalization of Gleditsia sinensis Lam.in China based on MaxEnt and ArcGIS.Lishizhen Medicine and Materia Medica Research 2022, 33(05):1201–1204.

19. Yin Qingfeng,Wang Yi,Huyan Li,Hao Jiabo,Meng Jinfang,Lu Bin.Genetic Diversity of Wild Zanthoxylum Armatum by ddRAD-seq.Molecular Plant Breeding.1–22·
20. Zhang AnShi,Luo Yang,Fan Ding-Chen,Zhang Zhong-Hai.Genetic diversity and fingerprints of Gleditsia sinensis germplasm resource based on SCoT.Guihaia 2017, 37(11):1378–1385.
21. Zhang Anshi,Zhang Sumin,Fan Dingchen,Liu Ying.Genetic diversity and fingerprints of Gleditsia sinensis germplasm based on SRAP.Acta Agriculturae Zhejiangensis2017, 29(09):1524–153

## Figures

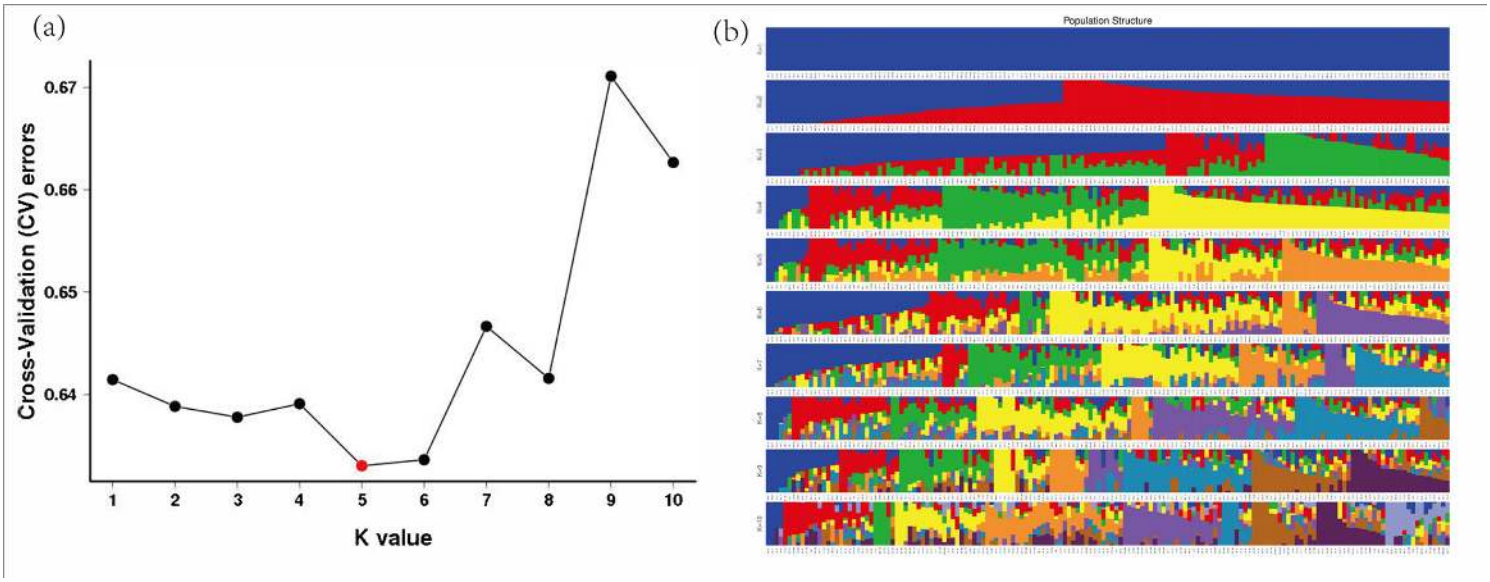
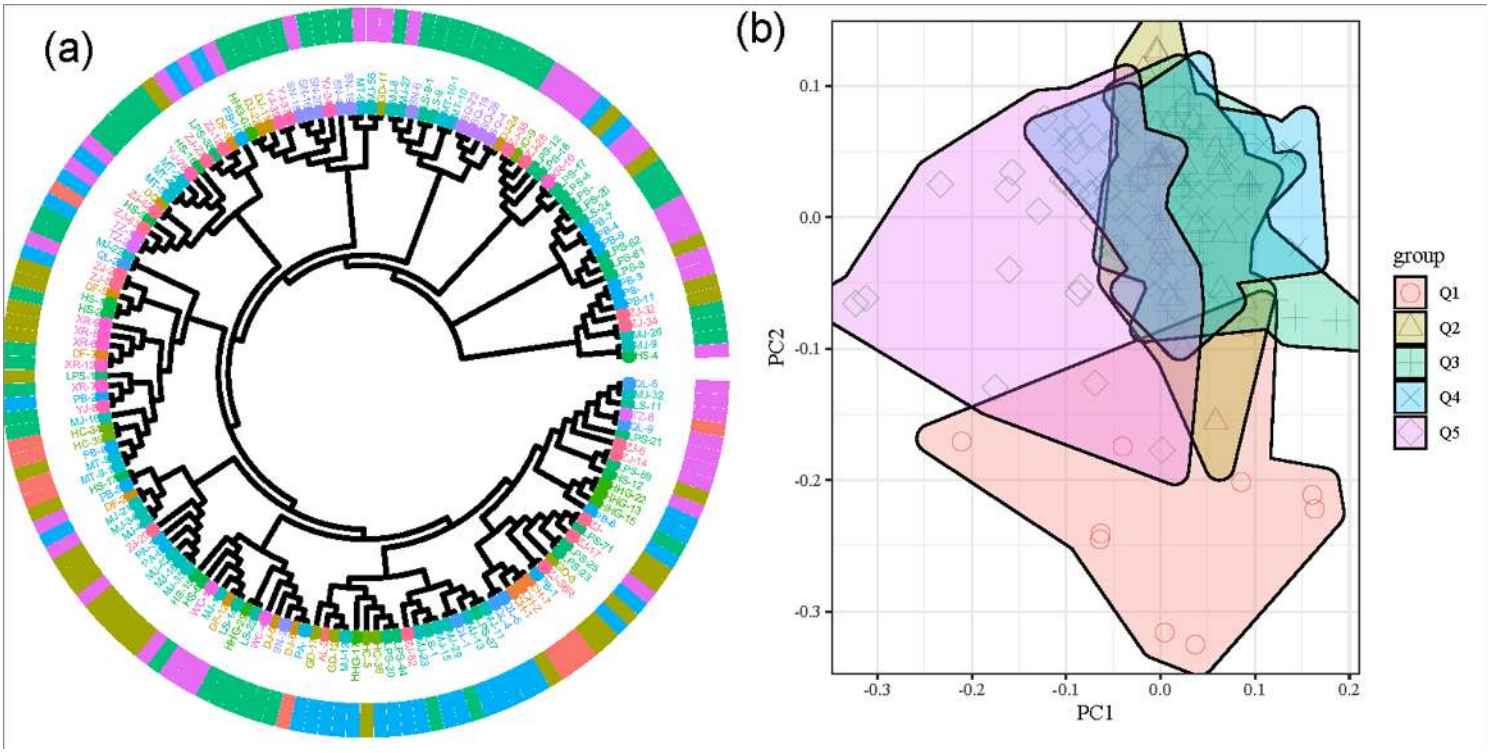



**Figure 1**

admixture validation error values (a) and Gentic structure cluster (b)

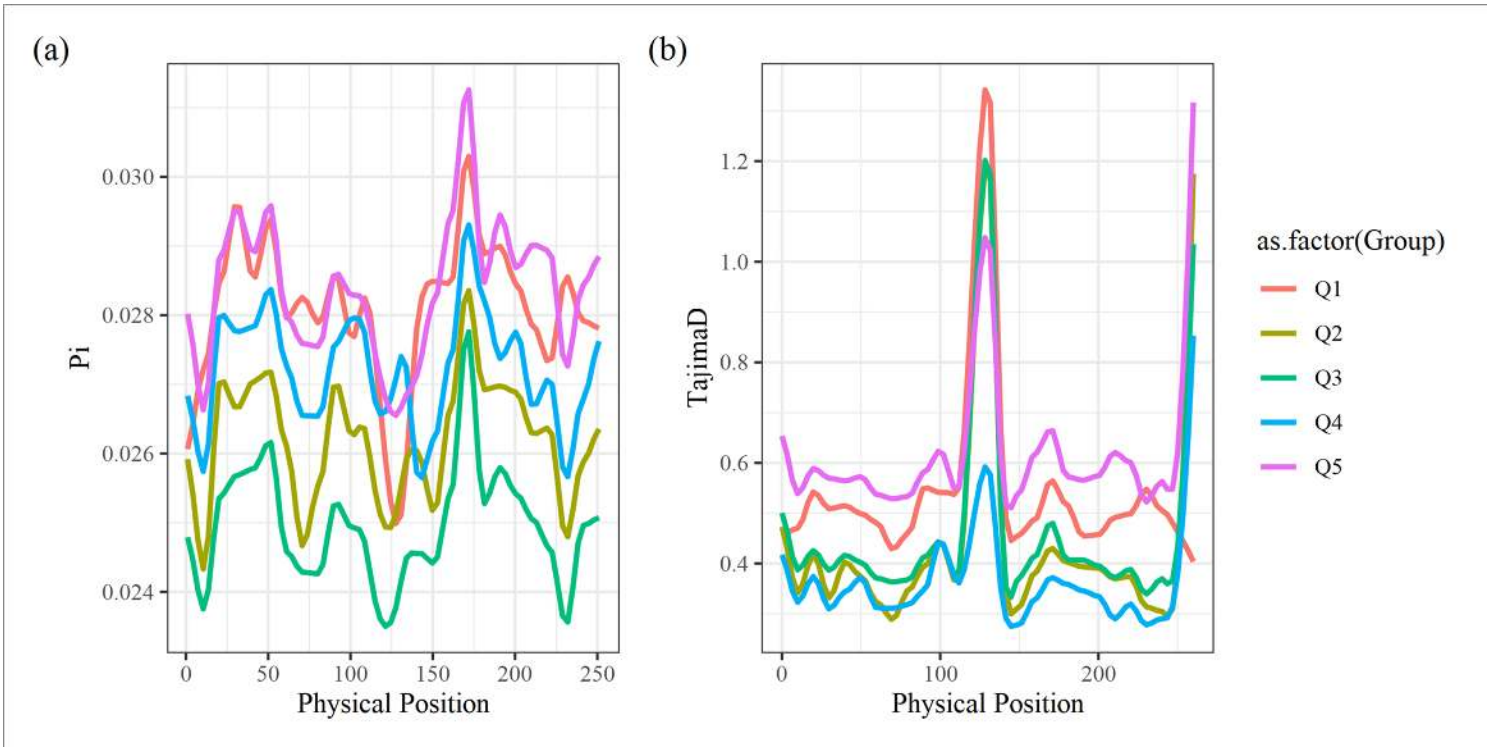


## Figure 2

Phylogenetic tree analysis and PCA analysis

(a) Phylogenetic tree analysis;(b) PCA analysis

Note: The names of the samples in the inner circle of Figure a were mapped according to administrative divisions, while the outer circle was mapped according to the population structure of admixtures; In Figure b, different populations were divided according to their population structure, and the periphery was drawn using geomer_mark'hull.




## Figure 3

Comparison of population diversity parameters

(a) The nucleotide diversity levels of different populations; (b) Comparison of Tajima's D values for different populations.

Note: The sliding window size was set to 10, and the curve fitting model is based on loess by the geom_smooth.