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.

Genetic Differentiation Analysis of Three Glycyrrhiza Species via Whole-Genome Resequencing

Jinmei Luo
Northwest University
Sani Zheng
Northwest University
Rongxing Liu
Northwest University
Jie Feng
Northwest University
Gang Zhou
Xinjiang Institute for Drug Control
Hong Tao
Xinjiang Institute for Drug Control
Jun Zhu
Xinjiang Uygur Autonomous Region Institute of Pharmaceutical Research
Zu'en Ji
Xinjiang Key Laboratory of Licorice and Its Products
Bin Guo
guobin@nwu.edu.cn
Northwest University
Yingjuan Wang
Northwest University

Research Article

Keywords: whole-genome resequencing, licorice, population evolution, genetic diversity

Posted Date: May 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9374564/v1

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

Additional Declarations: No competing interests reported.

Abstract

Glycyrrhiza species are leguminous plants of significant medicinal value.Growing recognition of their medicinal and commercial importance has led to a substantial increase in market demand.Despite this demand, the genetic differentiation and adaptive evolutionary mechanisms at the whole-genome level among core species such as *Glycyrrhiza uralensis*, *Glycyrrhiza glabra*, and *Glycyrrhiza inflata* remain poorly understood.This study collected 45 samples of these three licorice species from Xinjiang, China, and performed whole-genome sequencing using the Illumina HiSeq platform. Using approximately 4.08 Gb of high-quality data, we conducted single-nucleotide polymorphism (SNP) detection, population genetic structure analysis (including phylogenetic tree construction, principal component analysis, and population structure modeling), selective sweep analysis, and gene functional enrichment analysis to systematically elucidate the genetic relationships and adaptive differentiation among the three species. Over 10 million high-quality SNPs were identified.Population genetic analysis revealed three genetically differentiated groups at the genomic level. *Glycyrrhiza glabra* and *Glycyrrhiza inflata* were most closely related (Fst = 0.169), while both showed higher differentiation from *Glycyrrhiza uralensis* (Fst > 0.182). Selective sweep analysis identified genomic regions under strong natural selection during species differentiation. Functional enrichment analysis of candidate genes within these regions showed significant enrichment in biological processes and pathways such as flavonoid biosynthesis, cellular redox homeostasis, and transmembrane signaling—functions closely linked to licorice's secondary metabolism and environmental adaptability. This study demonstrates significant genetic differentiation among three medicinal licorice species at the whole-genome level and identifies potential functional genes associated with their adaptive evolution. These findings enhance our understanding of *Glycyrrhiza* evolutionary history and provide valuable genetic insights for medicinal germplasm improvement and resource conservation.

Introduction

*Licoric*e is a widely used medicinal plant, a perennial herb of the Fabaceae family, distributed across many arid and semi-arid regions worldwide[1]. This species is drought-tolerant and has a deep root system, playing a pivotal role in desert and semi-arid ecosystems. Genetic diversity is a crucial indicator of allele and genotype composition within populations and an important tool for investigating genetic relationships and evolutionary dynamics between populations[2]. Geographic isolation and long-term adaptation to different habitats have driven significant differentiation among populations in morphological traits, physiological responses, active component content[3], and genetic material[4]. Studies show that morphological variation in natural G. uralensis populations is primarily concentrated in leaves, which are highly susceptible to environmental influences.Leaf epidermal traits (e.g., stomatal size, density, and index) correlate significantly with geographical factors like light, temperature, and moisture[5]. In contrast, morphological variation in natural G. glabra and G. inflata populations is mainly reflected in pods and is influenced by temperature, photoperiod, pH, and moisture. Furthermore, key environmental constraints (e.g., water, light, and salt stress) can cause significant differences in physiological and biochemical indicators (e.g., biomass, photosynthetic pigment content, peroxidase and

superoxide dismutase activity) and secondary metabolite levels (e.g., glycyrrhizic acid, isoliquiritin, and liquiritin)[6,,[7]. DNA molecular studies have extensively confirmed rich genetic variation in natural G. uralensis populations.For example, Ge et al.[8]revealed extensive genetic variation within Chinese wild G. uralensis populations using amplified fragment length polymorphism (AFLP) markers, with variation significantly correlated with geographic distribution.Liu et al.[9]further demonstrated, using transcriptome-derived simple sequence repeat (SSR) markers, that genetic variation among natural G. uralensis populations exceeds interspecific levels.

In recent years, genomic research on Glycyrrhiza species has made initial progress, but existing work has primarily focused on single-species genome analysis, gene mining for specific metabolic pathways, or molecular marker development. For example, Zhang H et al.[10]analyzed simple sequence repeat (SSR) patterns based on whole-genome data from Glycyrrhiza uralensis; Dang H et al.[11]revealed the distribution patterns and genetic diversity of this species through resequencing populations of Glycyrrhiza uralensis from different habitats; Wu L et al.[12]conducted comparative and phylogenetic analyses of multiple licorice species at the organelle genome level by assembling chloroplast genomes. However, these studies remain limited in breadth and depth: At the whole-genome interspecific comparison level, particularly using population resequencing technology, no research has yet systematically analyzed the genetic differentiation, population history, and adaptive evolutionary mechanisms of the three core medicinal licorice species (Glycyrrhiza uralensis, Glycyrrhiza glabra, and Glycyrrhiza inflata).This research gap has adversely affected efforts to enhance the concentration of bioactive compounds and productivity in Glycyrrhiza species through molecular breeding. Previous analyses of genetic structure and diversity in licorice populations have primarily relied on microsatellite markers[9]and single nucleotide polymorphisms (SNPs)[9]. However, factors such as insufficient sample sizes and uneven marker distribution have limited comprehensive analyses of the whole-genome genetic diversity and associated regions in medicinal licorice species. To date, no studies have systematically elucidated the genetic differentiation and adaptive mechanisms of these three licorice species using whole-genome resequencing technology, highlighting the urgent need to explore this critical area.

Whole-genome resequencing in plants can effectively detect multiple DNA variation types, including SNPs, insertions/deletions (InDels), copy number variations (CNVs), and presence/absence variations (PAVs)[13[]–15], providing a foundation for in-depth analysis of genomic evolutionary mechanisms. WGS provides an in-depth view of an organism's complete genetic composition and is particularly effective for detecting genomic variations across different species[16]. Advances in sequencing technology have enabled widespread application of whole-genome sequencing (WGS) in multiple plant species, such as cotton breeding[17], Whole genome analysis of laccase gene family in wheat[18], and analysis of cassava domestication and agronomic traits[19]. The outcomes of these studies have broadened researchers' understanding of the application and understanding of genetic variation and adaptation in plants[20]. However, despite rapidly increasing medicinal and commercial demand for Glycyrrhiza species, genome-wide studies on the three major species (G. uralensis, G. glabra, G. inflata) remain limited.This gap hinders efforts to enhance bioactive compound concentration and productivity through molecular breeding.

Systematically revealing interspecific differences and genetic diversity at the whole-genome level among these three licorice species will provide new insights into the genetic mechanisms underlying their growth and development. Integrating whole-genome information for breeding design based on the genetics of phenotypic variation can meet target trait selection needs while maintaining sufficient genetic variation, enabling sustained genetic gains throughout the breeding cycle.

This study focuses on three licorice species from Xinjiang—G. uralensis, G. glabra, and G. inflata—using whole-genome resequencing (WGR) to analyze their genetic diversity, elucidate population genetic structures, and reveal adaptive evolutionary mechanisms.

Materials and Methods

1.1 Specimen Collection

This study collected three licorice species (15 samples each) from three locations in Xinjiang: G. uralensis from Toksun County, G. glabra from Wulan Township, and G. inflata from Yuli County.Based on Python/Basemap to generate a sample collection map The geographic distribution of the 45 samples is shown in Fig. S1. All 45 samples were subjected to whole-genome resequencing. In addition, all the samples collected in this study were identified by Professor Wenzhe Liu.

1.2 DNA Extraction and Genome Resequencing

Genomic DNA was extracted using the Plant Genomic DNA Extraction Kit (Tiangen Biotech, Beijing). DNA integrity and purity were assessed by 1% agarose gel electrophoresis.DNA purity was determined using a Nanodrop (OD260/280 ratio), and concentration was quantified with a Qubit 4.0 Fluorometer (Invitrogen). Sequencing libraries were prepared following the standard protocol of the Illumina TruSeq DNA PCR-free prep kit.

Specific steps were as follows: (1)DNA fragmentation and end repair: Extracted DNA was randomly fragmented by sonication, followed by end repair using the kit's End Repair Mix 2, which excised 5' overhangs, phosphorylated 5' ends, and filled in 3' recessed ends. (2)3'-end A-tailing: A single 'A' base was added to the 3' end of repaired DNA fragments to prevent self-ligation and facilitate ligation to sequencing adapters (with a 3' protruding 'T' base). (3)Adapter ligation: Sequencing adapters containing specific index sequences were ligated to the 5' ends of DNA fragments for subsequent immobilization on the Flow Cell. (4)Removal of adapter-self-ligated fragments: Ligation products were purified using BECKMAN AMPure XP Magnetic Beads to remove unligated or self-ligated fragments. (5)Library amplification and purification: Adapter-ligated DNA fragments were PCR-amplified to enrich the library. Amplification products were purified again with BECKMAN AMPure XP Magnetic Beads.(6)Library fragment selection: Libraries were size-selected and purified via 2% agarose gel electrophoresis, targeting a fragment size of approximately 450 bp.

Before sequencing, library quality was assessed using the Agilent High Sensitivity DNA Kit on the Agilent Bioanalyzer 2100 system.Qualified libraries showed a single major peak without adapter dimers. Library concentrations were precisely quantified using the Quant-iT PicoGreen dsDNA Assay Kit on the Promega QuantiFluor system, with a qualification threshold of ≥ 2 nM. Qualified libraries were gradient-diluted and pooled according to target sequencing volume ratios. Pooled libraries were denatured with NaOH to generate single-stranded DNA. Final sequencing was performed on the Illumina NovaSeq 6000 platform using paired-end 150-bp reads, achieving an average sequencing depth of 10×.

1.3 Sequencing Data Statistics, Quality Control, and Data Processing

This study simultaneously performed detection and screening of SNPs and InDels. To ensure the reliability of subsequent analyses, rigorous filtering of raw sequencing data is required to obtain high-quality data.

The data filtering workflow is as follows: (1)Remove adapter contamination: Use Adapter Removal (v2) software to remove adapter sequences from the 3' ends of reads. (2)Sliding window quality filtering: Base quality filtering was performed using a sliding window method (window size: 5 bp; step size: 1 bp). (3)Progressive window movement along the 5'→3' direction of reads: If the average quality score (Q) of bases within the current window ≤ 20, all sequences before the second-to-last base in the window were retained.If the Q value of the terminal base in the window ≤ 2, retain only the sequence preceding that base. (4)Length filtering: Remove any read ≤ 50 bp in length from the paired-end sequencing pair.

Use the bwa (0.7.12-r1039) mem program to align the filtered high-quality data to the reference genome, employing default bwamem parameters. Picard version 1.107 was used to sort the SAM file and convert it to a BAM file. The “FixMateInformation” command was employed to ensure consistency of paired-end read information.

GATK software was utilized for SNP detection, following these steps: (1)Realignment using known InDel information.This involved two steps: First, the Realigner Target Creator command within the GATK toolkit generated a file containing all potential InDels. Second, the IndelRealigner command was used to re-align reads near all InDels, enhancing SNP prediction accuracy. (2)The Unified Genotyper program was employed to identify SNP sites in samples, with stand-call-conf set to 30 and stand-emit-conf set to 10. Finally, SNP sites were annotated using the ANNOVAR software.

The UnifiedGenotyper program within the GenomeAnalysisTK v3.8 software package was used to identify all mutation sites in the samples, with stand-call-conf set to 30. The SelectVariants command was then employed to extract InDel mutation information. To ensure the reliability of InDel results, further filtering of InDel sites was performed using the following criteria: (1)Fisher Test of Strand Bias (FS) ≤ 200;(2)Read depth (DP) > 4; (3)Quality Depth (QD) ≥ 2; (4)ReadPosRankSum≥ -20;(5)Removal of sites with missing data.

1.4 Population Structure and Genetic Diversity Analysis

To elucidate evolutionary relationships among the three licorice populations, TreeBest 4 software was used to compute SNP-based genetic distance matrices. Phylogenetic trees were constructed using the neighbor-joining method, with branch support evaluated by 1000 bootstrap replicates. Principal component analysis (PCA) based on individual genomic SNP differences was performed to cluster individuals by genetic variation patterns. GCTA software computed eigenvalues and eigenvectors, visualized using R software.

Population genetic structure reflects non-random genetic variation distribution within species or populations, often manifesting as subpopulation differentiation due to geographic isolation. To reveal population structure and potential evolutionary history, the Bayesian model-based STRUCTURE software[21]inferred population genealogical information, with the optimal number of subpopulations (K) determined by the Evanno method[22].

Using selected SNP loci, Arlequin software[23]calculated genetic diversity indices for each population (parameters: minimum sequencing depth ≥ 4, missing rate < 0.1, minor allele frequency ((MAF) > 0.05). Observed heterozygosity (Ho) and expected heterozygosity (He) characterized within-population genetic variation levels, with high heterozygosity indicating greater genetic diversity.Nucleotide diversity (Pi) measured intrapopulation genetic diversity, reflecting average nucleotide variation among samples. The inbreeding coefficient (Fis) measured deviation of genotype frequencies from Hardy-Weinberg equilibrium within subpopulations. The population differentiation index (Fst) characterized genetic differentiation between subpopulations (range 0–1; lower Fst indicates less differentiation). Analysis of molecular variance (AMOVA) revealed sources of genetic variation among geographic populations. Identity-by-state (IBS) sequences denote identical allele fragments between individuals; IBS similarity reflects genetic similarity by calculating the proportion of shared allele sites.

1.5 Linkage Disequilibrium Analysis

Linkage disequilibrium (LD) refers to non-random association between alleles at different loci within a population.LD is characterized by the parameter R^2. LD analysis used VCFtools[24](v0.1.12b), assessing LD values across 10 kbp windows and calculating R^2 between SNPs separated by > 1 kbp.

1.6 Selective Clearance Analysis

The population fixation index (F) is a special case of Fst, reflecting allele heterozygosity levels within a population[24]. Sliding window analysis (window size 10 kb, step 20 kb) was performed on high-quality SNPs for selective sweep analysis. Combining Fst with SNP density effectively identifies regions under selection, particularly functional regions tied to survival environments that exhibit strong selection signals.Inter-population comparisons based on Fst and π aided target gene selection by screening for significant selection signals.

1.7 Gene Functional Enrichment Analysis

Genes within candidate regions identified by selective sweep analysis underwent Gene Ontology (GO) functional annotation and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment

analysis. In GO analysis, candidate genes were mapped to the GO database, and the number of genes associated with each functional term was counted.Hypergeometric distribution tests calculated enrichment significance (P-values), with Bonferroni correction ($P \leq 0.05$) identifying significantly enriched GO terms[25]. These terms revealed main biological processes, molecular functions, and cellular components involving candidate genes. KEGG is a comprehensive pathway database.Hypergeometric distribution tests similarly evaluated enrichment of candidate gene sets in specific pathways relative to the genomic background[26](corrected $P < 0.05$), identifying key metabolic and signaling pathways significantly enriched by candidate genes to elucidate their potential biological functions.

Results

2.1 Whole-Genome Resequencing of Three *Glycyrrhiza* Species

DNA extraction and whole-genome resequencing were performed on 45 licorice samples (15 per species), yielding 194.56 Gb of raw sequencing data. After quality control filtering, the average ambiguous base (N) proportion was 0.018%, and average GC content was 38.64%. The average proportions of bases with identification accuracy ≥ 99% (Q20) and ≥ 99.9% (Q30) were 97.29% and 93.64%, respectively.

Raw data contained adapter or low-quality reads potentially interfering with subsequent analysis. Strict filtering produced high-quality sequences. The basic statistics of the filtered data are as follows: high-quality reads comprised ~ 4.08 Gb; high-quality reads accounted for 97% of original reads and 94% of original base pairs; average sequencing depth was $10^{\times}$.

Filtered high-quality data (clean data) were aligned to the reference genome. Sample alignment rates ranged from 81.30% to 98.61%, with an average of 97%. Coverage analysis of reference genome sites (Fig. 1A-B) showed: average 82.86% of sites had coverage depth ≥ 1×; 64.01% had ≥ 4×; 23.31% had ≥ 10×; and 4.89% had ≥ 20×. Alignment rate reflects sample data similarity to the reference genome, while coverage depth and breadth indicate sequencing data uniformity and homology with the reference.Overall, alignment results were reliable for subsequent variant detection and related analyses.

2.2 Genome-Wide Variation Analysis (SNPs and Indels)

Based on whole-genome resequencing data from 45 licorice samples, we systematically identified two major types of genetic variation: single nucleotide polymorphisms (SNPs) and insertions/deletions (Indels). The number of SNPs on each chromosome was tallied, revealing the following results: Average homozygous genotypes consistent with the reference genome: 5,863,277-Average genotypes: 2,999,065-Average unknown genotypes: 738,133-Average homozygous genotypes inconsistent with the reference genome: 1,049,041 The average number of transpositions was 2,852,487; the average number of translocations was 1,195,619; and the average transposition-to-translocation ratio was approximately 2.390.

Chromosomal SNP density maps (Fig. 1C) were generated using SNPs identified in this study, visually illustrating their distribution across the genome. To investigate genomic distribution patterns, 10,649,516 sequenced SNPs were annotated (Table 1). Annotation results revealed: Among SNPs located in coding regions, 727,202 (6.83%) were in exons; 1,624,697 (15.26%) were in introns; and the highest number, 6,565,714 (61.65%), were in intergenic regions.

Using the same re-sequencing data, this study identified 2.35 million high-quality InDel sites. Their genomic distribution pattern resembled that of SNPs, with the vast majority located in non-coding regions. Among these, 53,107 InDels (2.26%) were found within exons.Notably, among these exonic InDels, frameshift indels—which can cause severe protein dysfunction—predominated, totaling 27,414.These high-impact functional variants, together with SNPs, constitute the genetic foundation driving species differentiation and adaptive evolution in licorice.

Table 1
Statistical Summary of SNP Annotation Results

Type	Number	Percentage
exonic total	727,202	6.83
synonymous SNV	316,336	2.97
nonsynonymous SNV	398,978	3.75
Stop gain	10,730	0.10
Stop loss	1,084	0.01
unknown	74	0.00
splicing	4,077	0.04
ncRNA total	0	0.00
ncRNA_exonic	0	0.00
ncRNA_splicing	0	0.00
ncRNA_exonic;splicing	0	0.00
ncRNA_intronic	0	0.00
intronic	1,624,697	15.26
intergenic	6,565,714	61.65
UTR5	0	0.00
UTR3	0	0.00
UTR5;UTR3	0	0.00
upstream	885,457	8.31
downstream	770,168	7.23
Upstream/downstream	72,201	0.68
Total	10,649,516	100

2.3 Population Structure and Genetic Diversity Analysis of Three *Glycyrrhiza* Species

Genetic diversity statistics for the three licorice populations (Table 2) showed significant differences in nucleotide diversity (Pi) between *G. uralensis* and the other two species (*G. inflata* and *G. glabra*), indicating higher genetic differentiation.In contrast, Pi values between *G. inflata* and *G. glabra* showed minimal variation, suggesting lower genetic divergence. Notably, inbreeding coefficients (Fis) for all three

populations were negative, indicating heterozygote excess and confirming none undergo self-fertilization.

To assess genetic differentiation among subpopulations, pairwise genetic differentiation indices (Fst) were calculated (Table 3) and visualized in a heatmap (Fig. 2A).Results showed: moderate genetic differentiation between *G. inflata* and *G. glabra* (Fst = 0.169166); high differentiation between *G. inflata* and *G. uralensis* (Fst = 0.293274); and moderate differentiation between *G. glabra* and G. uralensis (Fst = 0.182726).The heatmap visually represents pairwise genetic differentiation.

Table 2
Statistical Summary of Genetic Diversity for Each Population

Pop ID	Num Indv	Obs Het	Obs Hom	Exp Het	Exp Hom	Pi	Fis
Z	14.1834	0.2167	0.7833	0.1271	0.8729	0.1317	-0.1708
W	13.1026	0.3967	0.6033	0.3031	0.6969	0.3154	-0.1847
G	14.5962	0.2939	0.7061	0.1630	0.8370	0.1687	-0.2477

Table 3
Statistical Summary of Fst Results Among Species Populations

pop	Z	G	W
Z	-	0.169166	0.293274
G	-	-	0.182726

SNP-based phylogenetic analysis indicated that at the genomic level, the three licorice species formed three distinct branches, reflecting significant population differentiation (Fig. 2B). However, topological relationships still indicated some relatedness. Branch lengths revealed strong genetic differentiation among species, possibly due to reproductive isolation. PCA results showed the 45 individual samples grouped into three major clusters based on genomic SNP differences (Fig. 2C): red for *G. uralensis*, green for *G. inflata*, and blue for *G. glabra*. In the PC1-PC2 space, *G. uralensis* and *G. glabra*, as well as *G. uralensis* and *G. inflata*, were clearly distinguishable.However, some *G. glabra* and *G. inflata* individuals showed minor differences, resulting in partial clustering overlap, indicating heterogeneity in genetic distances among individuals across populations. SNP-based population genetic structure analysis revealed an optimal cluster number (K = 4), dividing all samples into four distinct subpopulations (Fig. 2D). *G. glabra* was split into two subpopulations, suggesting potential dual ancestral genetic origins. Significant genetic differentiation existed between *G. uralensis* and *G. inflata*, facilitating distinction. However, some *G. uralensis* and *G. glabra* individuals shared similar ancestral components, suggesting partial common ancestry.

2.4 Phylogenetic Analysis

To assess phylogenetic relationships, a phylogenetic matrix was constructed based on genome-wide SNP markers. Gmatrix (Ver2) software calculated the genomic relationship coefficient (G-value) between each pair of individuals (Fig. 3A). Results indicated *G. uralensis* was most distantly related to *G. inflata*, while *G. glabra* was relatively closer to *G. inflata*.

To further elucidate genetic similarity, plink (v1.9) software computed an identity-by-state (IBS) similarity matrix, converted to a genetic distance matrix (Fig. 3B). Analysis showed smaller genetic distance between *G. inflata* and *G. glabra* than with *G. uralensis*, suggesting a closer phylogenetic relationship, consistent with the G-matrix analysis. Notably, some *G. glabra* individuals did not cluster tightly, potentially due to factors like incomplete sharing of DNA sequence fragments, leading to relatively increased genetic distances.

2.5 Linkage Disequilibrium Analysis

Linkage disequilibrium (LD) analysis (Fig. S2) was quantified using the R^2 coefficient. Higher R^2 indicates stronger linkage between loci, typically observed at shorter physical distances between SNPs. Calculating the physical distance where LD decays to half its initial value (half-life distance) revealed that the *G. glabra* population exhibited high LD, potentially resulting from frequent inbreeding. *G. uralensis* and *G. inflata* populations showed moderate LD. Furthermore, all licorice populations had significantly higher LD decay rates within the 0–0.5 kb SNP interval, indicating large effective population sizes where allele frequencies are largely unaffected by genetic drift.

2.6 Selective Sweep Analysis

To assess population differentiation and identify differentiated genomic regions, inter-population fixation index (Fst) was calculated based on genome-wide SNP data (Fig. 4). Statistical analysis revealed: *G. glabra* vs *G. uralensis* (Fig. 4A): 1,320 strong selective sweep regions (mean $\theta\pi = 0.5499$, mean Fst = 0.2625), indicating significant genetic differentiation; *G. glabra* vs *G. inflata* (Fig. 4B): 1,006 strong selective sweep regions (mean $\theta\pi = 0.9191$, mean Fst = 0.2459), reflecting high genetic similarity; *G. uralensis* vs *G. inflata* (Fig. 4C): 858 strong selective sweep regions (mean $\theta\pi = 0.3709$, mean Fst = 0.3898), confirming low genetic similarity.

2.7 Gene Functional Enrichment Analysis

To elucidate biological functions of selected genes, functional enrichment analysis was conducted using Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases.

GO classifications include Molecular Function (MF), Biological Process (BP), and Cellular Component (CC). Fig. S3 shows comparison-specific enrichment patterns:

(1)*G. glabra* vs *G. uralensis* (Fig. S3A): Bubble chart: Intramolecular lyase activity had the most enriched genes.Bar chart: MF: Significant enrichment in beta-glucanase activity, glucan endo-1,3-beta-D-

glucosidase activity (C-3 substrate specificity), chalcone isomerase activity; BP:Flavonoid metabolic process and biosynthetic process were prominent.

(2)*G. glabra* vs *G. inflata* (Fig. S3B): Bubble chart:Generation of precursor metabolites and energy had the most enriched genes; Bar chart:MF: Transmembrane receptor protein kinase activity, transmembrane signaling receptor activity significantly enriched; BP: Cell redox homeostasis highest proportion; CC: Number of enriched terms significantly higher than G-vs-W.

(3)*G. uralensis* vs *G. inflata* (Fig. S3C): Bubble chart: Cellular process had the most enriched genes. Bar chart: MF: Chalcone isomerase activity dominated; BP: Flavonoid metabolic and biosynthetic processes most significant.

KEGG enrichment analysis of selected genes is shown via bubble plot (left) and bar chart (right) in Fig. S4:

(1)*G. glabra* vs *G. uralensis* (Fig. S4A): Bubble chart: Genes related to intramolecular lyase activity most significantly enriched; Bar chart: Enrichment primarily in Environmental Information Processing: Neuroactive ligand-receptor interaction; Metabolic Pathways: Fructose and mannose metabolism, Flavonoid biosynthesis; Organismal Systems.

(2)*G. glabra* vs *G. inflata* (Fig. S4B): Bubble chart: Biosynthesis of secondary metabolites dominated; Bar chart: Enrichment primarily in Organismal Systems: Toll and Imd signaling pathway; Metabolic Pathways:Phenylpropanoid biosynthesis; Genetic Information Processing: Sulfur relay system, Aminoacyl-tRNA biosynthesis.

(3)*G. uralensis* vs *G. inflata* (Fig. S4C): Bubble chart: Spliceosome-related genes most significantly enriched; Bar chart:Enrichment primarily in Metabolic pathways: Flavonoid biosynthesis; Genetic information processing: Nucleotide excision repair, ATP-dependent chromatin remodeling, Spliceosome; Human diseases: Human papillomavirus infection; Cellular processes: Cell cycle.

Discussion

Based on whole-genome resequencing data, this study identified three genetically distinct populations at the genomic level among the three *Glycyrrhiza* species. *Glycyrrhiza glabra* and *Glycyrrhiza inflata* showed close phylogenetic relationships, while *Glycyrrhiza uralensis* exhibited greater differentiation from the former two. This finding correlates with the chromosome staining study by Meng et al.[27], who demonstrated highly conserved chromosome structures among *Glycyrrhiza* species without significant interchromosomal rearrangements. This suggests interspecific differentiation in *Glycyrrhiza* likely relies more on sequence variation within gene regions or epigenetic regulatory mechanisms than on large-scale chromosomal structural changes.

This study performed whole-genome resequencing on 45 *Glycyrrhiza* samples to elucidate evolutionary patterns and genomic differentiation.After quality control, 4.08 GB of high-quality data and 10,649,516

high-confidence SNPs were obtained.

Exonic SNPs accounted for 6.83% (727,202), intronic for 15.26% (1,624,697), and intergenic for 61.65% (6,565,714). These mutations may drive phenotypic differentiation and ecological adaptation: Mutations in these regions significantly influence growth and development, linking licorice plant stress responses to bioactive compound synthesis[28–30]. He et al.[31]reported that the up-regulation of squalene synthase (SQS) and b-vanillin synthase (bAS) gave photogenic *Glycyrrhiza uralensis* seedlings and adult plants drought stress resistance, and the up-regulation of cytochrome P450 monooxygenase (CYP 450) gene was positively correlated with glycyrrhizic acid, a bioactive component of *Glycyrrhiza uralensis* Fisch. Gao et al.[32]demonstrated licorice extracts induce expression of nuclear factor erythroid 2-related factor 2 (Nrf2) and its downstream genes, which protect against toxic xenobiotics. Genomic analyses reveal that high genetic differentiation in *Glycyrrhiza* is a consequence of extensive ecological isolation, with populations distributed across a range from temperate steppes to desert regions[33]facing significant water gradients, intense light, high temperature fluctuations, and salt stress. This habitat heterogeneity may drive adaptive mutation accumulation through positive selection, providing a molecular basis for genetic improvement.

We employed whole-genome resequencing to investigate genetic structure and population evolution of three licorice species. Phylogenetic and principal component analyses revealed distinct genetic differentiation among Xinjiang *G. uralensis*, *G. glabra*, and *G. inflata* populations. Xinjiang's complex geography may have led to long-term, low-level distribution of *Glycyrrhiza*, with prolonged geographic isolation among the three populations.Geographic barriers likely caused differentiation and discontinuous distribution patterns.

In *Glycyrrhiza*, population genetic differentiation is primarily regulated by gene flow, as evidenced by highly differentiated populations resulting from its prolonged interruption, while hybrid groups act as genetic bridges to maintain connectivity[33]. Population structure analysis at optimal K = 4 divided all samples into four genetic clusters. *G. glabra* formed two independent clusters, indicating dual ancestral origins; a clear genetic boundary existed between *G. uralensis* and *G. inflata*, while some *G. uralensis* and *G. glabra* individuals shared ancestral components. The maximum likelihood phylogenetic tree confirmed no detectable gene flow among the 45 samples, indicating no significant population mixing events during their evolution. This study detected no significant population admixture events, suggesting that the three licorice species may have undergone independent genetic differentiation during their evolution. However, the natural triploid *G. glandulosa* discovered by Meng et al.[27]identified a natural triploid *G. glandulosa* in Xinjiang, demonstrating the evolutionary potential for new species formation through hybridization and polyploidy within *Glycyrrhiza*.Although polyploid samples were not included in this study, future research should focus on the distribution of polyploid resources within *Glycyrrhiza* and their contribution to genetic diversity and adaptive evolution.

Genetic diversity analysis revealed Ho values of 0.6969, 0.837, and 0.8729 for G. uralensis, *G. glabra*, and *G. inflata*, respectively; He values were 0.3031, 0.163, and 0.1271.Extremely low He values indicate low

genetic diversity levels in the three licorice species, potentially linked to habitat fragmentation and reduced effective population size. However, the unusually high Ho values deviate from expectations under classical small-population theory, suggesting strong selection pressures for heterozygote advantage and recent admixture between subpopulations of differing genetic backgrounds. This unique genetic pattern likely correlates closely with high-intensity habitat disturbance caused by human activities, though specific mechanisms warrant further investigation.

Based on whole-genome variation analysis, this study revealed genetic differentiation and evolutionary characteristics: *G. uralensis* had significantly higher nucleotide diversity (Pi) than *G. glabra* and *G. inflata* ($P < 0.01$), while Pi between *G. inflata* and *G. glabra* showed no significant difference ($P > 0.05$). Fixation index (Fst) analysis indicated moderate differentiation between *G. inflata* and *G. glabra* (Fst = 0.169166) and between *G. glabra* and *G. uralensis* (Fst = 0.182726), while G. inflata and *G. uralensis* showed high differentiation (Fst = 0.293274). Inbreeding coefficients (Fis) were negative (Fis < 0) for all three populations, indicating predominantly outcrossing. Linkage disequilibrium (LD) analysis further revealed: G. glabra exhibited high LD (potentially related to inbreeding), while other populations showed moderate LD. All populations showed rapid LD decay within 0–0.5 kb, consistent with the genetic diversity gradient patter[34, 35]: *G. glabra* > *G. inflata* > *G. uralensis*, suggesting *G. glabra* retained richer genetic variation during evolution.

Positive selection reduces intraspecific genetic diversity while intensifying interspecific differentiation. Under selective elimination, beneficial alleles driven by positive selection exhibit significantly elevated frequencies. As a core evolutionary driver, positive selection profoundly shapes genomic differentiation patterns in licorice populations, though background selection fails to fully suppress this differentiation process. It is particularly important to note that functional gene characterization in regions of low genetic diversity requires caution, as their variation patterns are difficult to assess accurately. Therefore, functional genomics research is urgently needed to elucidate these key genes and clarify the regulatory mechanisms of balancing selection in the adaptive evolution of licorice.

Bioactive compounds in medicinal plants, as products of gene-environment coevolution[36], have their synthesis pathways directly determined by herbal quality.Functional genomic analysis in this study revealed that selected genes in *Glycyrrhiza glabra* and *Glycyrrhiza uralensis* were significantly enriched in molecular fission enzyme activity (MF) and flavonoid metabolism (BP); *Glycyrrhiza uralensis* and *Glycyrrhiza inflata* showed higher abundance in transmembrane receptor kinase activity (MF), cellular redox homeostasis (BP), and cellular components (CC) compared to *Glycyrrhiza glabra*, indicating specific adaptations. In contrast, the *Glycyrrhiza uralensis-Glycyrrhiza inflata* comparison group exhibited prominent enrichment in cellular processes (BP). KEGG pathway analysis further revealed: neuroactive ligand-receptor interactions (*Glycyrrhiza glabra* vs. *Glycyrrhiza uralensis*) mediate environmental stress responses; the Toll/Imd signaling pathway (*Glycyrrhiza glabra* vs. *Glycyrrhiza inflata*) reflects evolutionary immune defense; flavonoid biosynthesis (*Glycyrrhiza uralensis* vs. *Glycyrrhiza inflata*) enhances desert habitat adaptation by scavenging reactive oxygen species (ROS).These naturally selected metabolic strategies provide a molecular framework for the ecological adaptation of *Glycyrrhiza*, offering guidance

for medicinal resource conservation and breeding. Through selective clearance analysis, this study revealed that the three licorice species underwent strong selection in pathways such as flavonoid synthesis and redox homeostasis, which are closely related to licorice's secondary metabolism.Notably, Duan et al.[37]proposed through ancestral state reconstruction that glycyrrhizin evolved independently twice within *Glycyrrhiza*: once in the broader *G. glabra* clade (including *G. uralensis*, *G. inflata*, etc.) and again in the North American species *G. lepidota*. Although this study did not directly measure glycyrrhizin content, its analysis of genomic selection signals provides insights into the genetic basis and evolutionary mechanisms underlying licorice's medicinal components.

Declarations

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. All authors confirm that the research was conducted without any external influences that might have biased the study design, data collection, analysis, interpretation, or the writing of the manuscript. The findings presented are based on objective research and analysis, and there are no conflicts of interest, whether financial, professional, or personal, that could potentially undermine the integrity of the study.

Data and resource availability

The datasets generated during and/or analysed during the current study are available in the NCBI repository, [PRJNA 1437893].

Funding

This work was supported by Autonomous Region Key Research and Development Program Project (No.2024B02024).

Author Contribution

Conceptualization: Jinmei Luo, Bin Guo; Methodology: Jinmei Luo, Sani Zheng, Xingrong Liu; Validation: Jinmei Luo, Sani Zheng, Xingrong Liu, Jie Feng; Formal analysis: Jinmei Luo, Sani Zheng, Jie Feng; Investigation: Jinmei Luo, Sani Zheng, Xingrong Liu, Jie Feng; Resources: Bin Guo, Gang Zhou; Writing-original draft: Jinmei Luo; Writing-review & editing: Sani Zheng, Bin Guo, Yingjuan Wang; Visualization: Sani Zheng; Supervision: Yingjuan Wang; Project administration: Bin Guo, Gang Zhou; Funding acquisition: Gang Zhou, Zu'en Ji, Hong Tao, Jun Zhu.

Acknowledgement

We thank the members of the Autonomous Region Key Research and Development Program Project (2024B02024) for their financial support of this study!

References

1. Liu Y, Li Y, Luo W, et al. Soil potassium is correlated with root secondary metabolites and root-associated core bacteria in licorice of different ages [J]. Plant and soil, 2020, 456(1–2): 61–79.
2. Muccillo L, Colantuoni V, Sciarrillo R, et al. Molecular and environmental analysis of Campania (Italy) sweet cherry (Prunus avium L.) cultivars for biocultural refugia identification and conservation [J]. Scientific reports, 2019, 9(1): 6796.
3. Yang R, Li W, Yuan B, et al. The genetic and chemical diversity in three original plants of licorice, Glycyrriza uralensis Fisch., *Glycyrrhiza inflata* Bat. and *Glycyrrhiza glabra* L [J]. Pakistan journal of pharmaceutical sciences, 2018, 31(2): 525–535.
4. Hosseini M S, EbrahimiI M, Samsampour D, et al. Association analysis and molecular tagging of phytochemicals in the endangered medicinal plant licorice (*Glycyrrhiza glabra* L.) [J]. Phytochemistry, 2021, 183: 112629.
5. Lu Ying L Y, Liu Renquan L R, Wang Wenquan W W, et al. Application of factor analysis to studying environmental factors with influencing effects on geographical distribution of *Glycyrrhiza* Radix et Rhizoma [J]. 2011: 1822–1827.
6. Li W D, Hou J L, Wang W Q, et al. Effect of water deficit on biomass production and accumulation of secondary metabolites in roots of *Glycyrrhiza uralensis* [J]. Russian journal of plant physiology: a comprehensive Russian journal on modern phytophysiology, 2011, 58(3): 538–542.
7. HosseiniI M S, Samsampour D, Ebrahimi M, et al. Study of Physiological and Biochemical Changes of Iraninan Licorice (*Glycyrrhiza Glabra*) under Salinity Stress in Filed Condition [J]. Journal of Crop Breeding, 2019: 193–201.
8. Ge S, Li G, Ma Z, et al. Analysis on genetic diversity of wild populations of licorice (*Glycyrrhiza uralensis* Fisch.) with AFLP markers [J]. Sci Agric Sin, 2009, 42(01): 47–54.
9. LIU Y, GENG Y, SONG M, et al. Genetic structure and diversity of *Glycyrrhiza* populations based on transcriptome SSR markers [J]. Plant Molecular Biology Reporter, 2019, 37(5): 401–412.
10. Zhan H, Wang Y, Du C, et al. SSR Molecular Markers Development Based on Whole Genome Sequences in Glycyrrhiza uralensis Fisch[J]. Medicinal Plant Research, 2020, 10(4): 1–8.
11. Dang H, Zhang T, Li Y, et al. Population evolution, genetic diversity and structure of the medicinal legume, *Glycyrrhiza uralensis* and the effects of geographical distribution on leaves nutrient elements and photosynthesis[J]. Frontiers in plant science, 2022, 12: 708709.
12. Wu L, Fan P, Cai J, et al. Comparative genomics and phylogenomics of the genus Glycyrrhiza (Fabaceae) based on chloroplast genomes[J]. Frontiers in Pharmacology, 2024, 15: 1371390.

13. Huang X, Kurata N, Wei X, et al. A map of rice genome variation reveals the origin of cultivated rice [J]. Nature, 2012, 490(7421): 497–501.
14. Jiao Y, Zhao H, Ren L, et al. Genome-wide genetic changes during modern breeding of maize [J]. Nature genetics, 2012, 44(7): 812–815.
15. Xue G-P, Mcintyre C L, Glassop D, et al. Use of expression analysis to dissect alterations in carbohydrate metabolism in wheat leaves during drought stress [J]. Plant Molecular Biology, 2008, 67(3): 197–214.
16. Lu Y, Li M, Gao Z, et al. Advances in whole genome sequencing: Methods, tools, and applications in population genomics[J]. International Journal of Molecular Sciences, 2025, 26(1): 372.
17. Wang P, Abbas M, He J, et al. Advances in genome sequencing and artificially induced mutation provides new avenues for cotton breeding[J]. Frontiers in Plant Science, 2024, 15: 1400201.
18. Zhong X, Li M, Zhang M, et al. Genome-wide analysis of the laccase gene family in wheat and relationship with arbuscular mycorrhizal colonization[J]. Planta, 2023, 257(1): 15.
19. Varshney R K, Saxena R K, Upadhyaya H D, et al. Whole-genome resequencing of 292 pigeonpea accessions identifies genomic regions associated with domestication and agronomic traits [J]. Nat Genet, 2017, 49(7): 1082–1088.
20. Dang H, Zhang T, Li Y, et al. Population evolution, genetic diversity and structure of the medicinal legume, *Glycyrrhiza uralensis* and the effects of geographical distribution on leaves nutrient elements and photosynthesis[J]. Frontiers in plant science, 2022, 12: 708709.
21. Pritchard J K, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data [J]. Genetics, 2000, 155(2): 945–959.
22. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study [J]. Molecular ecology, 2005, 14(8): 2611–2620.
23. Excoffier L, Laval G, Schneider S. Arlequin (version 3.0): an integrated software package for population genetics data analysis [J]. Evolutionary bioinformatics online, 2007, 1: 47–50.
24. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools [J]. Bioinformatics (Oxford, England), 2011, 27(15): 2156–2158.
25. Du Z, Zhou X, Ling Y, et al. agriGO: a GO analysis toolkit for the agricultural community [J]. Nucleic acids research, 2010, 38(Web Server issue): W64-70.
26. Mao X, Cai T, Olyarchuk J G, et al. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary [J]. Bioinformatics (Oxford, England), 2005, 21(19): 3787–3793.
27. Meng Z, Zheng Q, Shi S, et al. Whole-chromosome oligo-painting in licorice unveils interspecific chromosomal evolutionary relationships and possible origin of triploid genome species[J]. The Plant Journal, 2024, 120(5): 2089–2100.
28. Xanthopoulou A, Montero-Pau J, Mellidou I, et al. Whole-genome resequencing of Cucurbita pepo morphotypes to discover genomic variants associated with morphology and horticulturally valuable

traits [J]. Horticulture research, 2019, 6: 94.
29. Wang N, Zhang Y, Huang S, et al. Defect in Brnym1, a magnesium-dechelatase protein, causes a stay-green phenotype in an EMS-mutagenized Chinese cabbage (Brassica campestris L. ssp. pekinensis) line [J]. Horticulture research, 2020, 7: 8.
30. Nasrollahi V, Mirzaie-Asl A, Piri K, et al. The effect of drought stress on the expression of key genes involved in the biosynthesis of triterpenoid saponins in liquorice (*Glycyrrhiza glabra*) [J]. Phytochemistry, 2014, 103: 32–37.
31. He J, Yao L, Wang J, et al. Transcriptomic Analysis Reveals the Mechanism of Glycyrrhizic acid Biosynthesis in *Glycyrrhiza uralensis* Fisch [J]. 2021.
32. Gao N, Li Y, Zhang L, et al. The administration of *Glycyrrhiza* polysaccharides mitigates liver injury in mice caused by mancozeb via the Keap1-Nrf2/NF-κB pathway[J]. Food and Chemical Toxicology, 2025, 195: 115088.
33. Kim J, Lee J, Kang J S, et al. Contributions of interspecific hybrids to genetic variability in *Glycyrrhiza uralensis* and G. glabra[J]. Scientific Reports, 2025, 15(1): 8764.
34. Guo Y, Shen Y H, Sun W, et al. Nucleotide diversity and selection signature in the domesticated silkworm, Bombyx mori, and wild silkworm, Bombyx mandarina [J]. Journal of insect science (Online), 2011, 11: 155.
35. Xia J H, Bai Z, Meng Z, et al. Signatures of selection in tilapia revealed by whole genome resequencing [J]. Scientific reports, 2015, 5: 14168.
36. Li Y, Wu H. The research progress of the correlation between growth development and dynamic accumulation of the effective components in medicinal plants [J]. Chinese Bulletin of Botany, 2018, 53(3): 293–304.
37. Duan L, Han L N, Liu B B, et al. Species delimitation of the liquorice tribe (Leguminosae: Glycyrrhizeae) based on phylogenomic and machine learning analyses[J]. Journal of Systematics and Evolution, 2023, 61(1): 22–41.

Figures

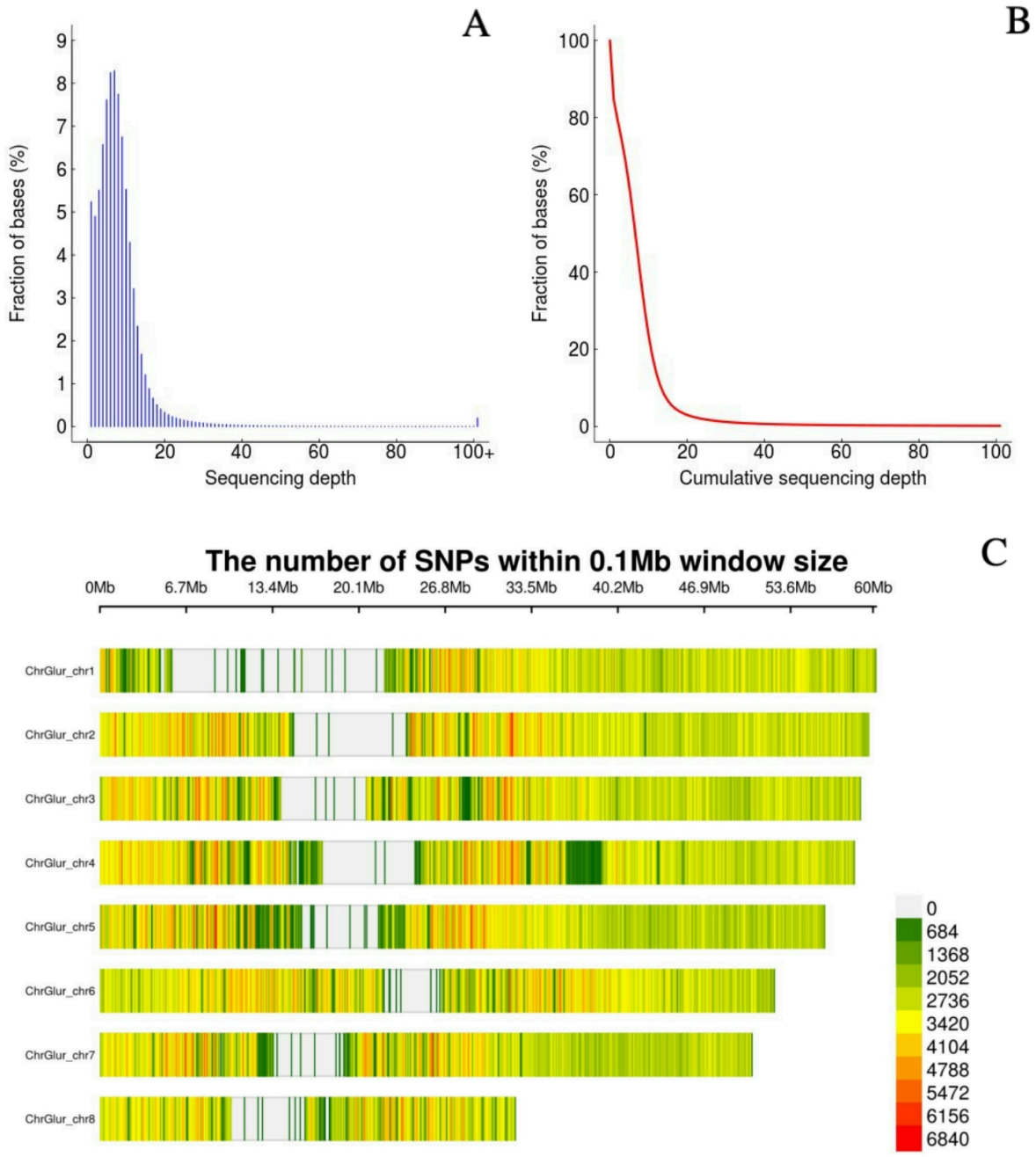


Figure 1

Sequencing Quality and SNP Distribution. (A-B)Sequencing depth and coverage. (C)Chromosomal distribution of SNPs.

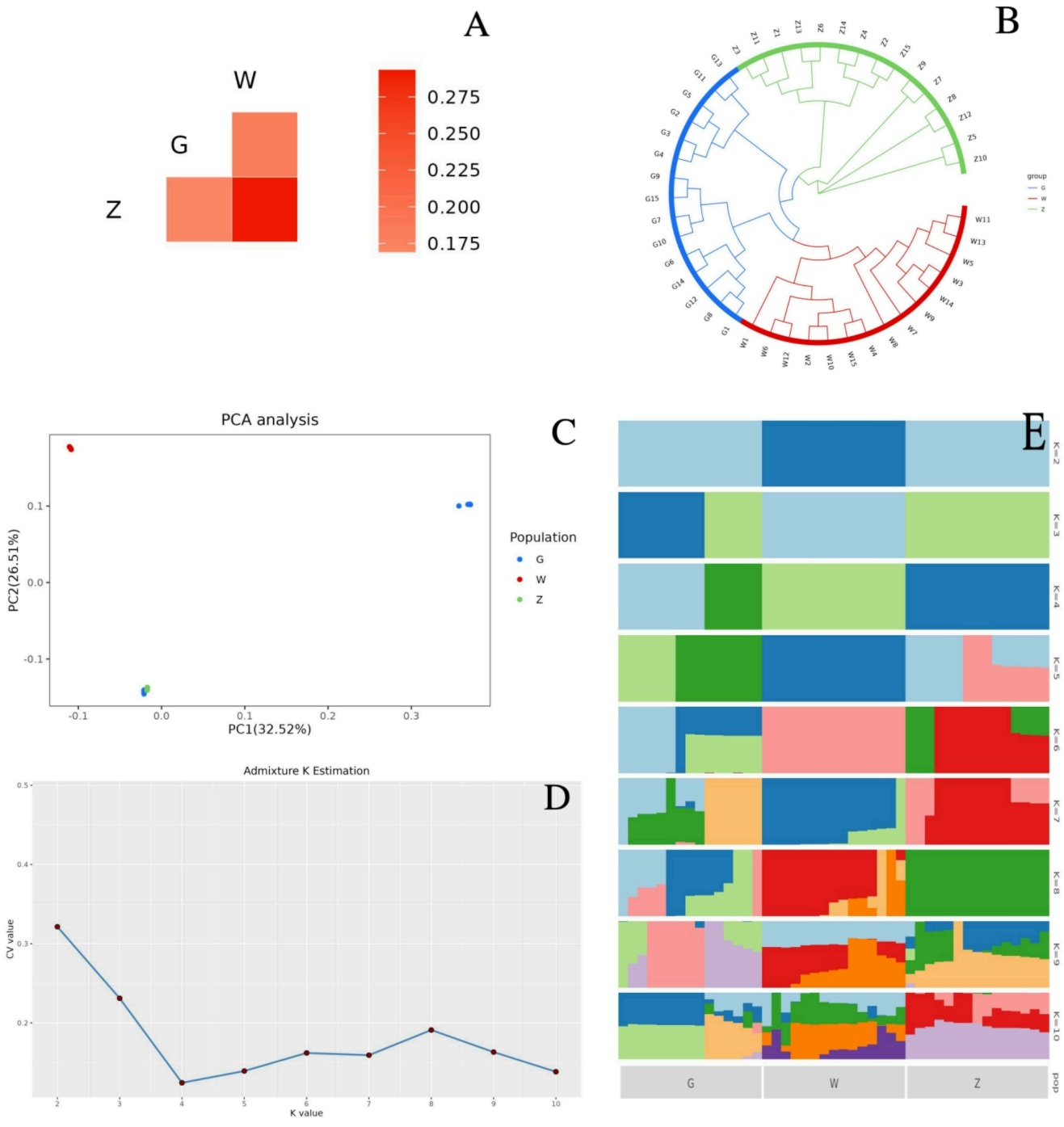


Figure 2

Population genetic structure analysis of three licorice species. Samples clustered into three distinct branches (*G. uralensis*-W, *G. glabra*-G, *G. inflata*-Z). (A)Population Fst heatmap; color intensity indicates differentiation level. (B)Phylogenetic tree showing significant genetic differentiation between species; branch lengths reflect genetic distance. (C)Principal Component Analysis (PCA) plot showing clear separation into three genetic clusters; points represent samples, colors indicate species, distances

reflect genetic background differences. (D)Cross-validation error at different K values; each matrix element represents the G value between corresponding individuals. (E)Genetic structure plot.

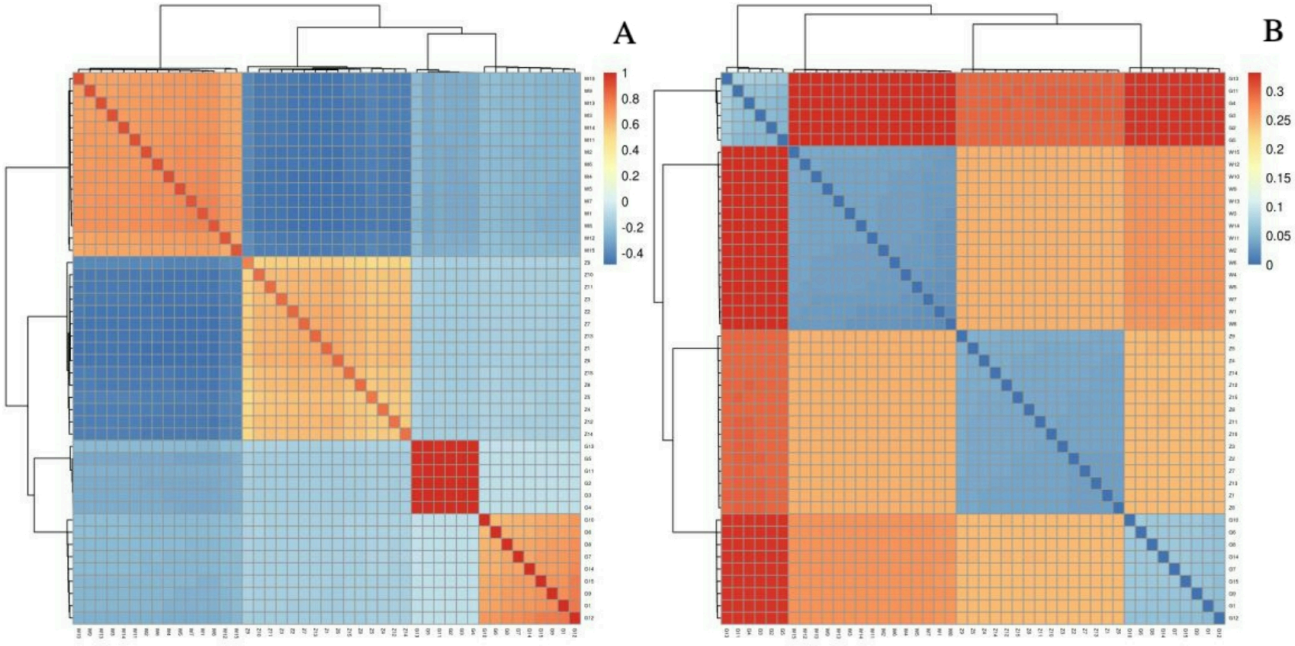


Figure 3

Phylogenetic Matrix and Linkage Disequilibrium. (A)Phylogenetic matrix; rows/columns correspond to sample IDs; color gradient indicates G-value intensity (red: high G-value, close relationship; blue: low G-value, distant relationship). (B)IBS genetic distance matrix; each element represents genetic distance between individuals; color gradient indicates distance magnitude (blue: small distance, close kinship; red: large distance, distant kinship).

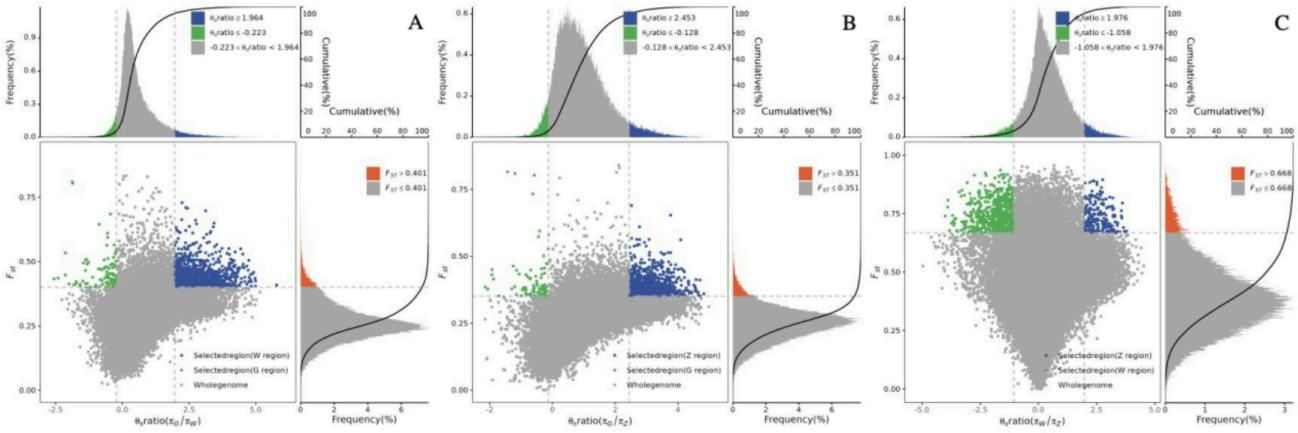


Figure 4

Selective sweep scatter plots. X-axis: θπ ratio (test/control); Y-axis: Fst values. Gray regions indicate neutral evolution; blue/green regions indicate selective sweeps. (A)*G. glabra* vs *G. uralensis*; (B)*G. glabra*vs *G. inflata*; (C)*G. uralensis*vs *G. inflata*.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryinformation.docx