scientific reports

OPEN

Contributions of interspecific hybrids to genetic variability in *Glycyrrhiza uralensis* and *G. glabra*

Jungeun Kim[1,11], Jeonghoon Lee[2,11], Jong-Soo Kang[3,10,11], Hyeonah Shim[3], Daewon Kang[3], Sae Hyun Lee[3], Jae-Pil Choi[2], Hui-Su Kim[4], Min Sun Kim[1], Yong Il Kim[2], Yunji Lee[2], Zafarjon Ziyaev[5], Yong Kook Shin[6]✉, Jong Bhak[1,4,7,8]✉ & Tae-Jin Yang[3,9]✉

Licorice (*Glycyrrhiza* L.), a medicinally and economically significant genus in the Fabaceae, is known for synthesizing glycyrrhizin. Here, we present a newly assembled genome of *Glycyrrhiza uralensis*, a key species distributed across Central and East Asia. Using Oxford Nanopore, Hi-C, and Illumina sequencing, we assembled a 415 Mbp genome with an N50 of 47 Mbp. While the genome structure was similar to previously reported assemblies, structural variations were observed across all eight chromosomes. Re-sequencing data from 29 individuals, including *G. uralensis*, *G. glabra*, and their hybrids, revealed significant genetic diversity, population structure, and hybridization events. Phylogenomic analyses using nuclear and plastid genomes demonstrated phylogenetic incongruence, supporting hybridization between *G. uralensis* and *G. glabra*. Our species network and gene flow test identified hybrid groups (Ggu1, Ggu2, and Gug) acting as genetic bridges between the species. Demographic history inferred via PSMC showed *Glycyrrhiza* species thrived during the Middle Pleistocene, with population size fluctuations in *G. uralensis* and hybrids. Current low heterozygosity and high genetic differentiation suggest long-term geographic and ecological isolation, reducing gene flow. Our findings advance the understanding of evolutionary history in *Glycyrrhiza* species and help conservation and molecular breeding of these species.

Licorice (*Glycyrrhiza* L.) is a plant widely utilized for medicinal and culinary purposes. In East Asia, including Korea and China, it has been traditionally used as a medicinal herb, while in Europe and the Middle East, it is popular as an ingredient in soft drinks, food, and snacks[1]. One of the most notable beneficial compounds in licorice is glycyrrhizin (glycyrrhizic acid), a type of triterpenoid saponin known for its physiological activity in treating neuroactive, inflammatory, allergic, arthritic, and cancer diseases[2–5]. Another valuable compound found in licorice is liquiritigenin, a type of flavanone that is pharmacologically utilized for its anti-inflammatory, antioxidant, anti-bacterial, and anti-diabetic effects[6–8].

The genus *Glycyrrhiza*, belonging to the Fabaceae family, comprises approximately 20 species of perennial herbs or subshrubs that are distributed worldwide, ranging from the Mediterranean to Mongolia, Australia, and the Americas[9]. Among the *Glycyrrhiza* species, five species biosynthesize glycyrrhizin: *G. uralensis*, *G. glabra*, *G. inflata*, *G. aspera*, and *G. lepidota*. Of these, *G. lepidota* is found only in North America, while the other four species are distributed across the Eurasian continent[9–12]. *G. uralensis* is primarily distributed from Central Asia to Northeast China; *G. glabra* from Europe to Northwest China (Xinjiang); *G. inflata* in Central Asia, Northwest China, and Mongolia; and *G. aspera* from Mongolia through Central Asia to Europe[8,9]. Additionally, *G. korshinskyi*, first described as a new species by Grigorjev[13], is known to have originated from

[1]Personal Genomics Institute, Genome Research Foundation, Cheongju 28190, Republic of Korea. [2]National Institute of Horticultural and Herbal Science, RDA, 27709 Eumseong, Republic of Korea. [3]Department of Agriculture, Forestry and Bioresources, College of Agriculture & Life Sciences, Plant Genomics & Breeding Institute, Research Institute of Agriculture and Life Science, Seoul National University, Seoul 08826, Republic of Korea. [4]Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Republic of Korea. [5]Research Institute of Plant Genetic Resources, Tashkent 111208, Uzbekistan. [6]School of Industrial Bio-pharmaceutical Science, Semyung University, 65 Semyung-ro, Jechon 27136, Chungcheonbuk-do, Republic of Korea. [7]Department of Biomedical Engineering, School of Life Sciences UNIST, Ulsan, Republic of Korea. [8]Clinomics Inc, Ulsan 44919, Republic of Korea. [9]Institutes of Green Bioscience and Technology, Seoul National University, Gangwon-do 25354, Republic of Korea. [10]Department of Forest Resources, College of Forest and Environmental Sciences, Kangwon National University, Chuncheon 24341, Republic of Korea. [11]Jungeun Kim, Jeonghoon Lee and Jong-Soo Kang contributed equally to this work. ✉email: shinella612@gmail.com; jongbhak@genomics.org; tjyang@snu.ac.kr

hybridization between *G. uralensis* and *G. glabra*. This species is found in Central Asia, including Xinjiang in China, Kazakhstan, and Kyrgyzstan[14].

The first draft genome of the glycyrrhizin-synthesizing species was published at scaffold level from *G. uralensis* by incorporating Illumina Hiseq-2000 short reads and PacBio RSII long reads[15], and then the *G. uralensis* genome was improved by Rai et al.[16] at chromosome level, utilizing a combination of HiFi and Hi-C sequencing reads. Previous studies on the genetic diversity of the glycyrrhizin-synthesizing species have been conducted at the population level using microsatellite or single nucleotide variant (SNV) data for *G. uralensis*, *G. glabra* and *G. inflata*[17–19]. Additionally, a study utilizing microsatellite markers has been carried out for all four species[20], but did not include interspecific hybrids or investigate the genetic structure and diversity of these hybrid individuals. Consequently, the genetic contribution of these hybrids within the species complex remains unexplored.

In this study, we aim to assemble a chromosome-level reference genome of *G. uralensis* using nanopore long reads, Hi-C sequencing, and Illumina short reads. We further utilize re-sequencing data from 29 *Glycyrrhiza* individuals collected from seven countries—China, Korea, Uzbekistan, Mongolia, Kazakhstan, Kyrgyzstan, and Russia—to elucidate the genetic structure and inheritance of *G. uralensis*, *G. glabra*, and their hybrids. Our study will provide a foundation for the conservation of genetic diversity and the molecular breeding of licorice, a valuable medicinal plant.

Results

Chromosome level assembly of *Glycyrrhiza uralensis* (CN01) collinear to the other licorice reference genome

G. uralensis (CN01) collected from China was used for whole genome sequencing (Supplementary Table S1) to obtain a high-quality reference genome assembly. With 145 M cleaned-up reads, the genome size of CN01 was estimated to be 380–400 Mb (Supplementary Fig. S1). The estimated genome size of CN01 was similar to two previously reported *G. uralensis* genomes of 391.9 and 405 Mbp in size[15,16]. A total of 16 million long-reads (56 GB) were generated by Oxford Nanopore PromethION and assembled by wtdbg2, resulting in 6826 contigs being assembled. Although the quality of our Hi-C data was not optimal, these contigs were then ordered into eight chromosomes using Hi-C contact information (Supplementary Table S1; Supplementary Fig. S2). The CN01 genome was eventually assembled into a total length of 415 Mbp with an N50 value of 47 Mbp. We identified 227 M repetitive sequences in the CN01 genome, representing 54.57% of the genome (Supplementary Table S2), which was higher than the first draft *G. uralensis* genome (39.44%) reported by Mochida et al.[15] and slightly lower than the one (61.7%) reported by Rai et al.[16]. The most abundant repeats in the CN01 genome were transposable elements, including retrotransposons (16.45%) and DNA transposons (9.26%) (Supplementary Table S2).

We predicted 35,422 protein-coding genes in the assembly. BUSCO values showed that 97.7% of the genes were identified with respect to the Fabales genomes, indicating high quality (Table 1). Syntenic analysis was conducted between the eight chromosomes of the CN01 assembly and the chromosome-level assembly of Rai et al.[16]. As a result, all eight chromosomes corresponded in a 1:1 manner with those in the previous assembly (Fig. 1a), with no significant structural differences observed (Supplementary Fig. S3). Structural variation analysis was conducted across the eight chromosomes of two *G. uralensis* genomes assembled to the chromosome level (Supplementary Table S3; Supplementary Fig. S4). This analysis indicated that structural variations, both greater than and less than 50 bp, were not concentrated on any specific chromosome but were evenly distributed across the eight chromosomes (Supplementary Fig. S4).

Comparative analyses with other closely related legume genomes

To investigate the function of duplicated genes in the CN01 genome, we performed gene expansion analyses among five legume genomes in the subfamily Papilionoideae: *Arachis duranensis* (tribe Dalbergieae), *Lotus japonicus* (tribe Loteae), *G. uralensis* (CN01), *Cicer arietinum*, and *Medicago truncatula* in the IR-lacking clade (Fig. 1b, c,d). Among the five legume genomes, 23,446 orthologous groups and 4,617 species-specific genes were identified by OrthoFinder (Fig. 1b). A total of 357 orthologous groups were duplicated in the common ancestor of the IR-lacking clade (*G. uralensis*, *C. arietinum*, and *M. truncatula*), and 2,708 orthologous groups were duplicated specifically in the *G. uralensis* genome (Fig. 1c). The gene ontology (GO) analysis indicated that the duplicated genes specifically in the *G. uralensis* genome were associated with ethylene signaling (Fig. 1d), which is related to plant growth[21].

	Mochida et al.[15]	Rai et al.[16]	CN01 (this study)
Total length (bp)	378,856,821 bp	459,214,121 bp	415,300,097 bp
No. of scaffolds/chromosomes	12,528 scaffolds	8 chromosomes	8 chromosomes
N50 (bp)	109,269	58,560,175	46,964,546
GC content (%)	35.34%	37.13%	34.98%
No. of protein coding genes	34,445	32,941	35,422
BUSCO (Fabales odb10)	97.6%	98.6%	97.7%

Table 1. Comparison of the assembled *Glycyrrhiza uralensis* genome compared with two previously reported genomes.

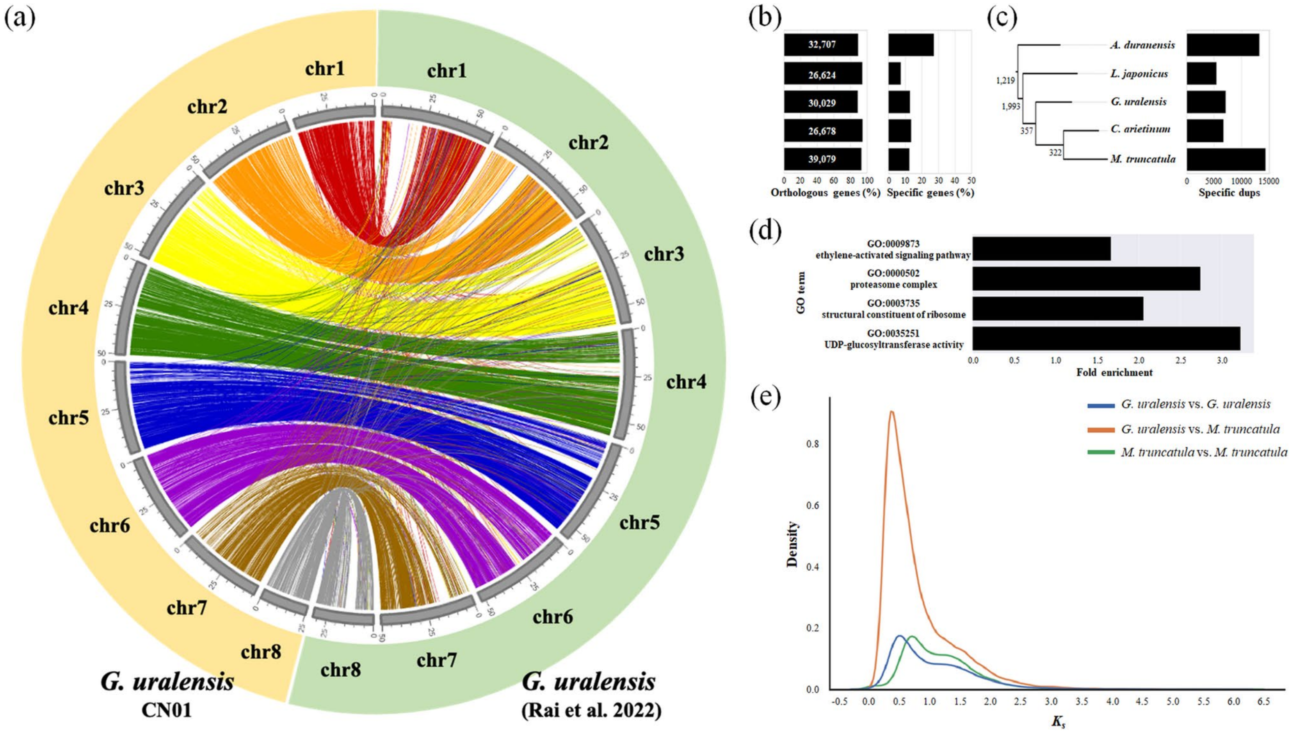

Fig. 1. Synteny and evolutionary analyses of *G. uralensis* with related species. (**a**) Synteny analysis between the newly assembled *G. uralensis* genome (CN01) and the previously reported *G. uralensis* genome[16]. (**b**) Distribution of synonymous substitution rates (Ks) between *G. uralensis*-*G. uralensis*, *G. uralensis*- *Medicago truncatula*, and *M. truncatula*-*M. truncatula* collinear gene pairs. (**c**) percentages of orthologous genes and species-specific genes in the *Arachis duranensis*, *Lotus japonicus*, *G. uralensis*, *Cicer arietinum*, and *Medicago truncatula* genomes. (**d**) Phylogenetic relationship of the five legume genomes and number of species-specific duplicated genes. The numbers near the branches represent number of duplications on each branch. (**e**) gene ontology (GO) enrichment analysis of the duplicated genes specifically in the *G. uralensis* genome.

To elucidate the evolutionary relationships between *G. uralensis* and three other legumes, we analyzed the distribution of synonymous substitutions per synonymous site (Ks) using collinear gene pairs (Fig. 1e, Supplementary Fig. S5). The Ks distribution between *A. duranensis*, which belongs to an early-diverging clade within the subfamily Papilionoideae, and the three species of the IR-lacking clade revealed two distinct peaks (Supplementary Fig. S5). The first peak, around a Ks value of 0.6, likely represented genetic differentiation between *A. duranensis* and the IR-lacking clade. The second peak, near a Ks value of 1.4, is inferred to correspond to a whole-genome duplication (WGD) event that occurred in the common ancestor of the IR-lacking clade (Supplementary Fig. S5). Additionally, in the Ks distribution of collinear gene pairs between the *G. uralensis* and *M. truncatula* genomes, a peak around the Ks value of 0.4 indicated genetic differentiation between these two species (Fig. 1e). Furthermore, in the Ks distribution of paralogous genes within each genome, *G. uralensis* showed a peak around Ks = 0.5, while *M. truncatula* showed a peak around Ks = 0.6, suggesting that these species share a WGD event. This finding supports the presence of a WGD event in the common ancestor of the IR-lacking clade, as inferred from comparisons with *A. duranensis* (Fig. 1e).

Phylogenetic analyses based on single-copy nuclear genes and plastid genomes

We performed phylogenetic analyses using both the concatenated method and the coalescent method with the selected 459 single-copy orthologous genes. In the phylogenetic analysis, with *G. pallidiflora* (KR05) as the outgroup, five *G. glabra* (Gg) individuals (UZ01, UZ02, CN06, CN07, and KG03) and four *G. uralensis* (Gu) individuals (CN01, CN02, CN03, and CN05) were placed in separate clades, which were consistent between the concatenated and coalescent trees (Supplementary Fig. S6). However, the phylogenetic positions of several *G. korshinskyi* individuals (MN03, RU01, KR01, KR02, and KR04) showed phylogenetic incongruence between the trees generated by the two methods. Additionally, while the support values in the concatenated tree were robust across almost all branches, the local posterior probability calculated by ASTRAL for the coalescent tree showed low supporting values, with many branches below 50% (Supplementary Fig. S6).

To infer maternal lineages, we assembled the plastid genomes (plastomes) of 29 *Glycyrrhiza* individuals (Supplementary Table S4). The assembled plastomes showed no genome rearrangements and exhibited the typical plastome structure of the IR-lacking clade (Supplementary Fig. S7), which is characterized by the loss of one copy of the Inverted Repeat (IR), consistent with previously reported genome structures[12]. Our plastome-based phylogeny appeared mostly similar to the coalescent tree constructed from 459 single-copy nuclear genes (Supplementary Fig. S6b), but key differences were observed (Fig. 2a). Our plastome-based phylogeny revealed

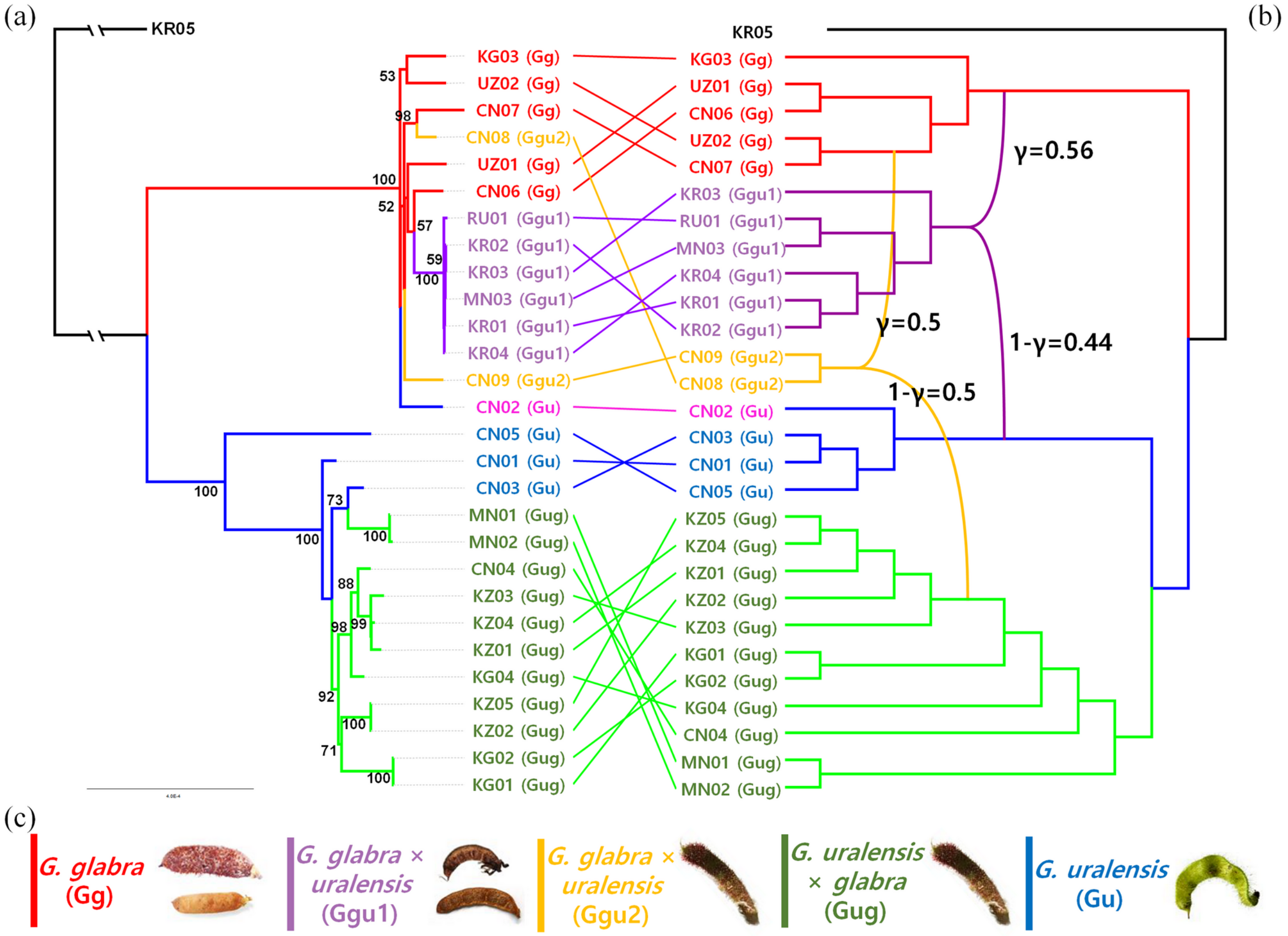

Fig. 2. Species network inference from single-copy nuclear genes and plastome-based phylogeny of 29 *Glycyrrhiza* individuals. *G. pallidiflora* (KR05) was included as the outgroup. (**a**) The plastome-based phylogeny was constructed using IQ-TREE. The values near the nodes represent bootstrap support values from the ML analysis. One *G. uralensis* individual (CN02), which has a *G. glabra*-type plastome, is indicated in pink. (**b**) The species network was inferred from 459 single-copy gene trees using PhyloNet. The γ and 1-γ values indicate inheritance probabilities from each parental clade. (**c**) Pod morphology of each group, with groups indicated by different colors in the plastome-based phylogeny (**a**) and species network inference (**b**).

two major clades, *G. uralensis* (Gu) and *G. glabra* (Gg), with strong support (100%). However, the hybrid species, *G. glabra* × *uralensis* groups (Ggu1 and Ggu2) and *G. uralensis* × *glabra* group (Gug), tend to co-cluster according to their maternal lineage. Both Ggu1 and Ggu2 formed a sister clade to Gg, while Gug formed a sister clade to Gu. This topology suggests that the phylogeny largely reflects the maternal genetic structure, even though one *G. uralensis* individual appears within the *G. glabra* clade (Fig. 2a).

Comparison of the plastome and nuclear coalescent trees revealed a similar topology; however, a few hybrids indicated more complicated relationships by exhibiting incongruences between phylogenetic positions inferred by maternal plastome and biparental nuclear genome data. Further analysis of the nuclear concatenated and coalescent trees revealed more complex patterns. Ggu1 was monophyletic in both the plastome phylogeny and the nuclear coalescent tree (Fig. 2a, Supplementary Fig. S6b), but it was not monophyletic in the nuclear concatenated tree (Supplementary Fig. S6a). This incongruence suggests the influence of biparental nuclear genome variation. The Ggu2 group, consisting of two individuals (CN08 and CN09), possessed a Gg-type plastome and clustered with Gg and Ggu1 in the plastome phylogeny (Fig. 2a). However, nuclear data revealed divergence: CN08 grouped as a sister to Gg, while CN09 formed a clade with Gug (Supplementary Fig. S6). This indicates potential back-crossing and gene flow between hybrid lineages and their parental species.

Among the hybrid groups, Gug showed the most stable phylogenetic position, consistently forming a clade with Gu across all analyses. Nonetheless, it was only monophyletic in the nuclear coalescent tree (Supplementary Fig. S6b). The consistent maternal lineage pattern of Gug, combined with its occasional non-monophyly in nuclear analyses, further suggests that hybridization and back-crossing events play a significant role in shaping the genetic structure of *Glycyrrhiza* species. Overall, these results highlight the complexity of hybridization in *Glycyrrhiza* and suggest that both maternal and biparental genetic data are necessary to fully understand the evolutionary relationships and hybridization dynamics in this genus.

Species network inference and interspecific gene flow test

We used PhyloNet to infer species networks and HyDe to perform gene flow tests to determine whether these groups were formed through hybridization and identify the parental lineages involved. Among the inferred species networks, all five hypotheses, which allowed for one reticulation, commonly inferred that the six individuals of the Ggu1 group (KR01, KR02, KR03, KR04, RU01, and MN03) and CN08 of the Ggu2 group were formed through hybridization between Gg and Gu (Supplementary Fig. S8a). Furthermore, in all five hypotheses allowing up to three reticulations, it was inferred that the six individuals of the Ggu2 group (CN08 and CN09) were likely formed through hybridization between two Gg individuals (UZ02 and CN07) and five individuals collected from Kazakhstan (KZ) of the Gug group (Fig. 2b, Supplementary Fig. S8b). The estimated inheritance probabilities (γ) indicated that the genetic compositions of the two Ggu groups (Ggu1 and Ggu2) were approximately equally derived from each parental lineage (Fig. 2b). These inferences suggest that licorice genetic diversity continues to increase by ongoing genetic hybridization between species, as well as between hybrids, in Central Asia.

To detect gene flow among the five defined groups, we conducted an interspecific gene flow analysis using HyDe. Using *G. pallidiflora* (KR05) as the outgroup and analyzing 35,588,300 SNVs for gene flow, hybridization signals were detected in the Ggu1, Ggu2, and Gug groups (Fig. 3a). The Ggu1 group was confirmed to result from hybridization between Gg and Gu (Z-score = 56.2999, P-value ≈ 0.0, γ = 0.5946). The Ggu2 group was found to have formed from hybridization between Gg and the Gug group (Z-score = 45.5679, P-value ≈ 0.0, γ = 0.2444), while the Gug group was determined to have originated from hybridization between Gg and Gu (Z-score = 10.0743, P-value ≈ 0.0, γ = 0.5927). In 1000 bootstrap replicates, hybridization was detected in Ggu1 and Ggu2 with 100% confidence, respectively, and in Gug with 92.9% confidence. The estimated γ values are presented in (Fig. 3a). The gene flow detected in Ggu1 and Ggu2 is consistent with the species network inferred by PhyloNet (Fig. 2b), while the hybrid origin of the Gug group was only detected by HyDe (Fig. 3a). The gene flow analysis showed that the γ (gamma) value for the Ggu1 group is consistently around 0.5, with minimal deviation. This indicates that these are likely F1 hybrids, with 4 out of 6 individuals being artificially bred between Gg and Gu in Korea. In contrast, the Gug group exhibited a wide range of γ values, suggesting that these hybrids formed naturally over a long period in the wild, with ongoing gene flow between Gg and Gu.

Principal component analysis and population structure

To infer the genetic structure of *Glycyrrhiza* species, a total of 36 M SNVs were identified from 482 Gb Illumina short reads of 29 *Glycyrrhiza* individuals (Supplementary Table S1). After stringent filtrations, approximately 3 M SNVs used for a subsequent principal component analysis (PCA) and admixture analysis. Using 3,130,455SNVs, PCA was performed (Fig. 3b). The PCA results revealed clear separation among the six groups: Gp, Gg, Gu, Ggu1, Ggu2, and Gug. *Glycyrrhiza glabra* (Gg) and *G. uralensis* (Gu) were mainly separated by PC1, and Ggu1 and Gug, which are inferred to have formed through gene flow between Gg and Gu, were primarily separated by PC1 (Fig. 3b). In the two dimensional space of PC1 and PC2, Ggu1 appeared closer to Gu, whereas Gug appeared closer to Gg. The two individuals of Ggu2 (CN08 and CN09) were positioned near Gug, with CN08 closer to Gg and CN09 closer to Gug (Fig. 3b).

We analyzed the genetic structure of 29 *Glycyrrhiza* individuals using the SNV data used in PCA analysis. The optimal value for K (i.e., the number of clusters) was determined to be 4, with the lowest cross-validation value. At K = 2, Gg and the Kazakhstan (KZ) and Kyrgyzstan (KG) individuals of the Gug group clustered in one group, and Gu in another, whereas Ggu1, Ggu2 and MN01, MN02, and CN04 of the Gug group appeared to be an admixed group (Fig. 3c). At K = 3, Gu remained as one group, UZ01 and CN06 of Gg formed a distinct group with KG01 and KG02 of the Gug, and the other KG and KZ individuals of the Gug group and UZ02 and CN07 of Gg clustered in the last group. The remaining four individuals (MN01, MN02, CN04, and KG04) of the Gug group, Ggu1, Ggu2, and KG03 of Gg appeared to be admixed groups, respectively. The Ggu1 group was admixed between Gg and Gu, and the four individuals of the Gug (MN01, MN02, CN04, and KG04) were admixed between Gu and Gug, and Ggu2 was between Gg and Gug (Fig. 3c). At K = 4, the structure was similar to that at K = 3, but within the Gug group, KG01 and KG02 remained as one group, while the remaining individuals in the Gug group appeared to be an admixed group. CN08 and CN09 were identified clearly as admixtures between Gg and Gug (Fig. 3c).

Morphological diversity in the *Glycyrrhiza* species

We compared the morphology of the leaf, inflorescence, flower, and pod (fruit) among *G. glabra* (Gg), *G. uralensis* (Gu), *G. glabra* × *uralensis* (Ggu), and *G. uralensis* × *glabra* (Gug) (Supplementary Table S6). Although the inflorescence of all *Glycyrrhiza* species was racemose, the shape of the inflorescence slightly differed. The Gu and Gg exhibited oval and oblong racemes, respectively, while Ggu and Gug showed a variety of inflorescence shapes ranging from oval to oblong, but their inflorescences were slightly closer to Gu (Supplementary Fig. S9; Supplementary Table S6). The size of the flowers in Gu and Gug was observed to be slightly larger than those in Gg and Ggu (Supplementary Fig. S9). Quantitative characters, including leaf length, inflorescence length, and pod length, showed that, unlike Gu, which has a short inflorescence, Ggu and Gug possess a long inflorescence similar to Gg (Supplementary Table S6). Additionally, these hybrid groups do not exhibit the curved ring-shaped pod characteristic of Gu, but they have longer pods compared to Gg (Supplementary Table S6). The pod morphology with trichomes is an important characteristic for distinguishing *Glycyrrhiza* species[22]. The Gu had curved ring-shaped pods with dense trichomes, while Gg had straight-growing pods, but the trichome density varied by region (Supplementary Fig. S9). The Gg individual (UZ02) collected from western Uzbekistan, near Europe, mostly had glabrous pods, whereas the Gg individual (UZ01) from eastern Uzbekistan, near Asia, had pods with high trichome density. The Ggu group had semi-ring shaped pods with trichomes, resembling Gu,

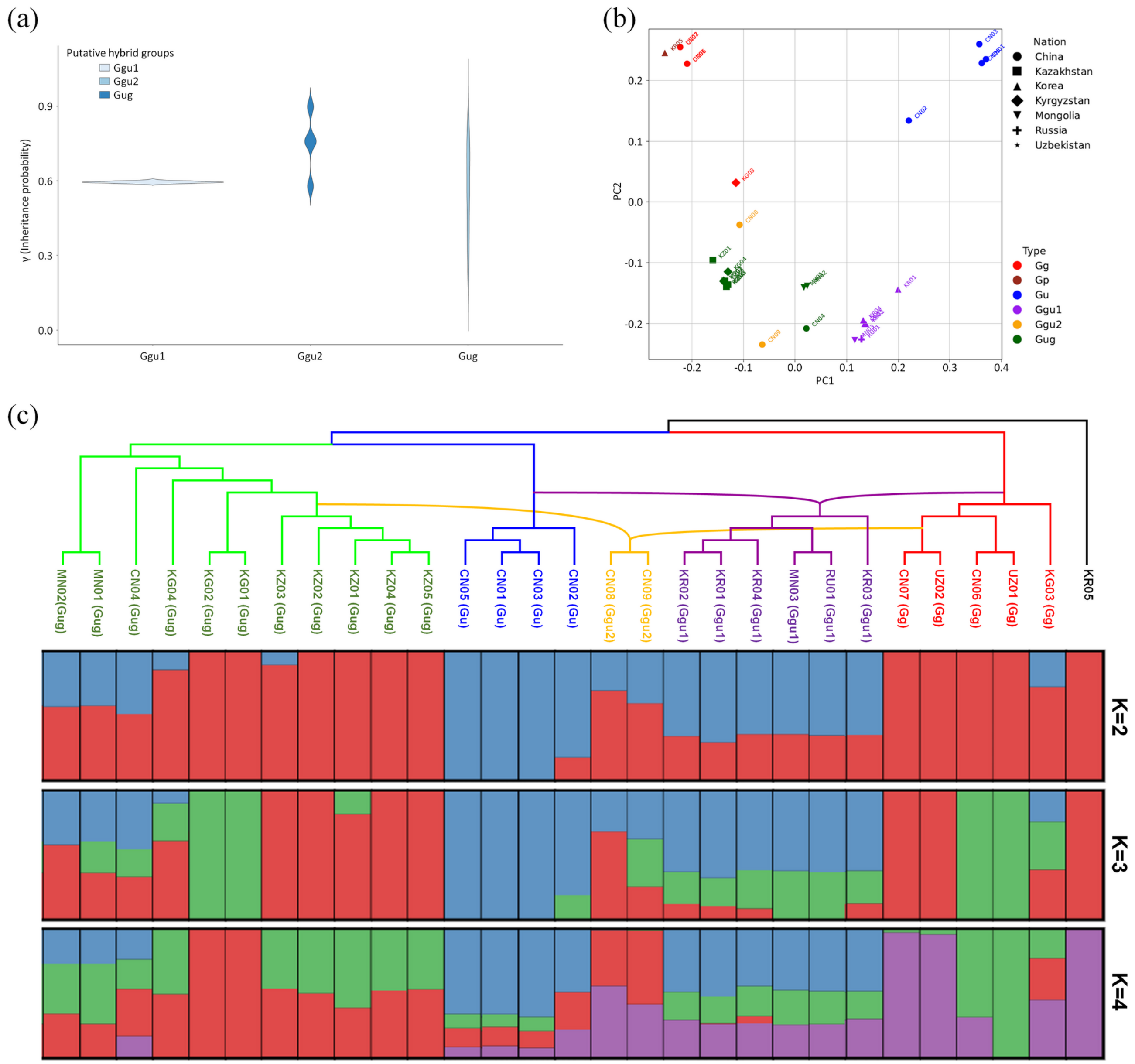

Fig. 3. Population structure and interspecific gene flow in *Glycyrrhiza* species. (**a**) Violin plot showing the distribution of γ values calculated by the interspecific gene flow test using HyDe across 1000 bootstrap replicates. Ggu1 and Gug were inferred as the hybrid origin groups between Gg and Gu. Ggu2 was inferred as the hybrid origin group between Gg and Gug. The γ and 1-γ values indicate the inheritance probabilities from the putative parental groups, respectively. (**b**) Principal component analysis (PCA) of 29 *Glycyrrhiza* individuals. Each nation is represented by a different symbol and each species or group is indicated by different colors. Gp: *G. pallidiflora*. (**c**) Population structure of 29 *Glycyrrhiza* individuals inferred using ADMIXTURE for K = 2 to K = 4. The results are arranged according to the species network inferred by PhyloNet. *Gu G. uralensis*, *Gg G. glabra*, *Ggu G. glabra* × *uralensis*, *Gug G. uralensis* × *glabra*.

while the Gug group had straight pods with trichomes, resembling Gg, showing an intermediate morphology between Gu and Gg, further supporting its hybrid origin (Supplementary Fig. S9; Supplementary Table S6).

Population history of licorice

To assess genetic diversity within the *Glycyrrhiza* populations, we calculated the heterozygosity of 29 *Glycyrrhiza* individuals (Fig. 4a). All 29 individuals exhibited heterozygosity below 0.01, with KR01, an individual formed through artificial hybridization between Gg and Gu[23], showing the highest heterozygosity (approximately 0.009), whereas *G. pallidiflora* (KR05) displayed the lowest heterozygosity (approximately 0.001). Between Gg and Gu, Gu exhibited slightly higher heterozygosity. The hybrid groups, particularly Ggu1 and Gug, showed relatively higher heterozygosity compared to the parental species, Gg and Gu (Fig. 4a). We also calculated the pairwise

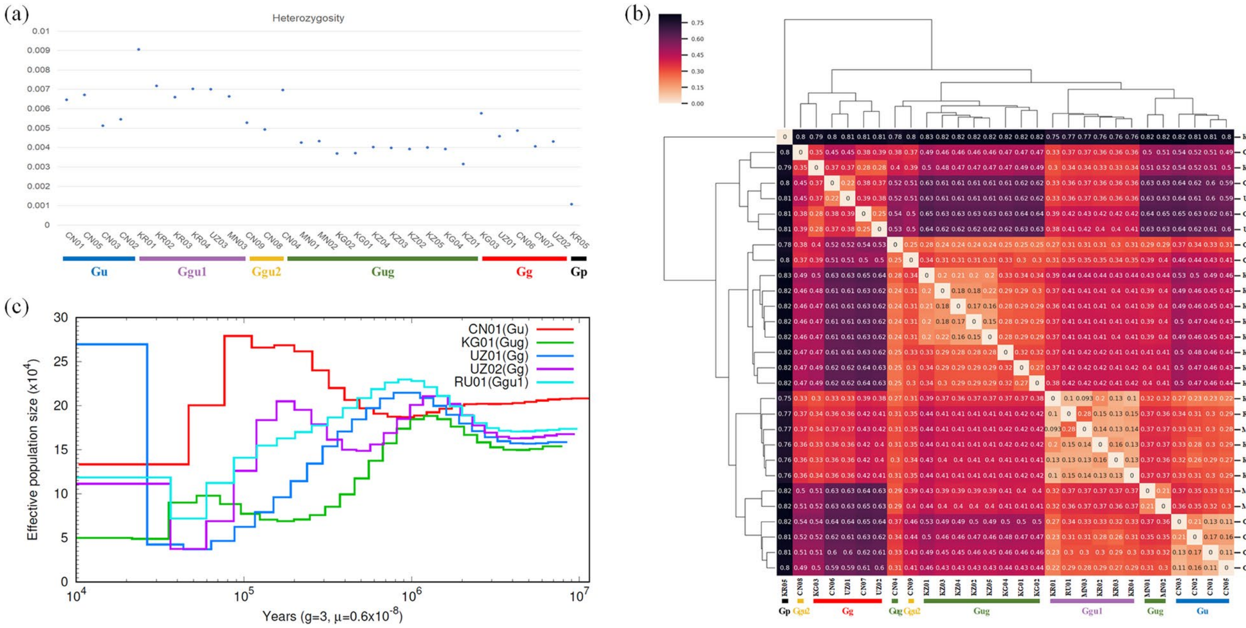

Fig. 4. Genomic diversity and population dynamics of *Glycyrrhiza* species. (**a**) Heterozygosity of 29 *Glycyrrhiza* individuals. The graph displays the variability in heterozygosity across 29 individuals. (**b**) Heatmap of fixation index (Fst) values among the 29 individuals. The heatmap represents selection with colors ranging from bright (less differentiated) to dark (highly differentiated). The heatmap was drawn using the Pong library in Python. (**c**) Historical effective population size changes for four representative *Glycyrrhiza* individuals. The effective population size for each individual was analyzed using the PSMC model. The variables "g" and "μ" represent generation time and mutation rate, respectively. *Gp G. pallidiflora*, *Gu G. uralensis*, *Gg G. glabra*, *Ggu G. glabra* × *uralensis*, *Gug G. uralensis* × *glabra*.

fixation index (F_{ST}) to estimate the degree of genetic differentiation among *Glycyrrhiza* species (Fig. 4b). The F_{ST} values between *G. pallidiflora* (KR05) and the other 28 *Glycyrrhiza* individuals, including Gg, Gu, Ggu1, Ggu2, and Gug, ranged from 0.75 to 0.83, indicating significantly high genetic differentiation. Within the 28 *Glycyrrhiza* individuals, Gg and both Gu and Gug showed the highest genetic differentiation by showing the F_{ST} values over 0.6. The Ggu1 group, formed through equal gene flow between Gg and Gu, had F_{ST} values of approximately 0.4 or lower (Fig. 4b). In the Ggu2 group, CN08 exhibited higher genetic differentiation with Gu compared to Gg, while CN09 showed higher differentiation with Gg compared to Gu, resulting in contrasting outcomes. Although the F_{ST} values within each group were lower than those between groups, they were still above 0.1 (Fig. 4b), suggesting a relatively high level of genetic differentiation and potential isolation within the *Glycyrrhiza* populations.

We used pairwise sequentially Markovian coalescent (PSMC) analysis to investigate the demographic history of *G. uralensis* and its related species by analyzing the effective population size (N_e). We examined N_e over the last 10 million years (Ma) for five representative individuals: CN01 from the Gu group, UZ01 and UZ02 from the Gg group, KG01 from the Gug group, and RU01 from the Ggu1 group (Fig. 4c). Although the N_e of Gu was higher than that of other Glycyrrhiza species, it tended to decreased between 10 Ma and 1 Ma (Fig. 4c). Meanwhile the N_e of Gu, Ggu1 and Gug increased during this period. Starting from 2 Ma, these *Glycyrrhiza* species exhibited dramatic changes in N_e. The N_e of Gu, Ggu1, Gug, and Gg showed an increase. Notably, the N_e of Gu and UZ02 (Gg) dramatically increases around 300 Ka, peaking at approximately 100 Ka. The N_e fluctuation of *Glycyrrhiza* over time provides insights into the hybridization among *Glycyrrhiza* species driven by ancient climate changes in natural conditions, as well as the distribution of modern *Glycyrrhiza* species and their hybrids (Fig. 4c).

Discussion

Although two *G. uralensis* genomes have been previously reported[15,16], we documented a newly assembled genome from a different accession of *G. uralensis* (CN01), which is a valuable medicinal resource. The first draft genome focused on elucidating the biosynthetic pathway of glycyrrhizin, a key pharmacological compound in *Glycyrrhiza* species[15]. Later, Rai et al.[16] focused on achieving a high-quality chromosome-level assembly of the *G. uralensis* genome. Our study also assembled a high quality of the *G. uralensis* (CN01) genome, which showed a higher level of completeness compared to that reported by Rai et al.[16] (Table 1; Fig. 1a, Supplementary Fig. S3). Furthermore, we identified substantial structural variations across the eight chromosomes (Supplementary Table S3; Supplementary Fig. S4). Although a reference genome has already been reported, reporting new genomes from other accessions aligns with the current trends in plant genomics. This approach is consistent with the concept of pan-genome studies, which address the genetic content and diversity that cannot be fully covered by a single reference genome by utilizing large-scale resources[24–28]. Therefore, the CN01 genome reported in this

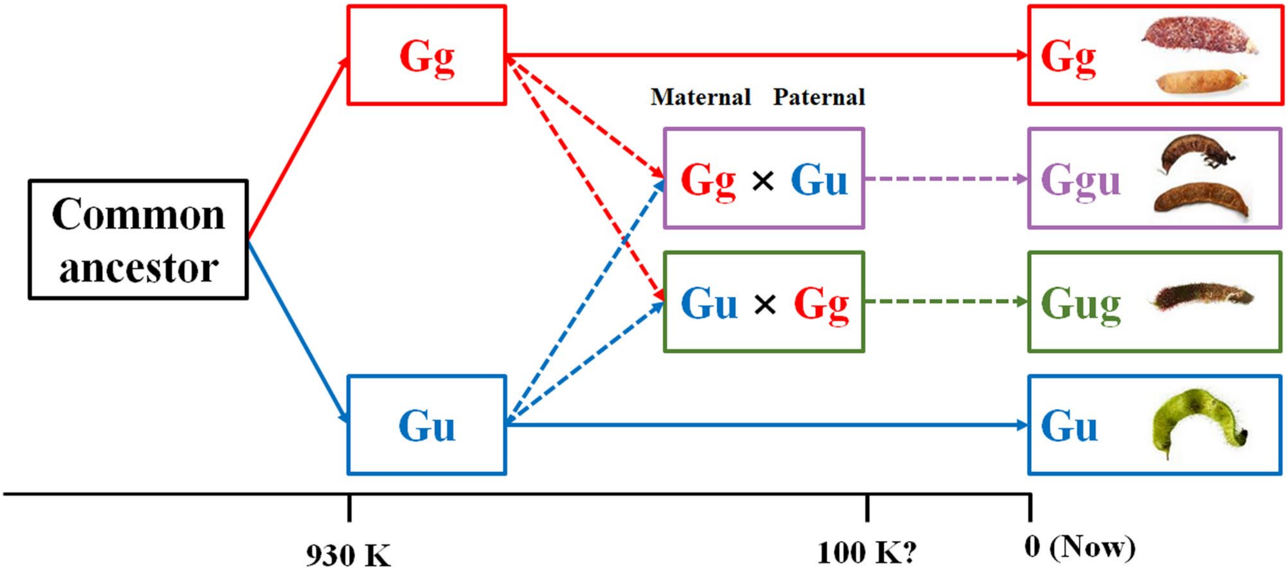

Fig. 5. Schematic representation of gene flow and pod morphology in *Glycyrrhiza* species. The schematic diagram summarizes the inferred evolutionary history and gene flow between *G. glabra* (Gg) and *G. uralensis* (Gu) based on the results of this study. Approximately 930,000 years ago (930 K), Gg and Gu diverged from a common ancestor[12]. Around 100,000 years ago (100 K), during a period of increased effective population size in *Glycyrrhiza* species as suggested by PSMC analysis, extensive gene flow occurred between the two species. This gene flow resulted in the formation of the Ggu group, with Gg as the maternal parent and Gu as the paternal parent, and the Gug group, with Gu and Gg as the maternal and paternal parents, respectively. The diagram also shows the pod morphology of the current Gg, Ggu, Gug, and Gu groups, illustrating their phenotypic differences.

study, which exhibited sufficient structural variations compared to the previously reported genomes, will serve as a crucial resource for future studies on genomics and molecular breeding of *G. uralensis* and its related species.

Based on re-sequencing data of 29 *Glycyrrhiza* individuals, this study focused on understanding the genetic diversity, genetic composition of hybrids, genetic differentiation, and demographic history of *Glycyrrhiza* species (Figs. 2, 3 and 4). From Central and East Asia, we collected a total of 29 *Glycyrrhiza* individuals, including *G. uralensis*, *G. glabra*, and their hybrid individuals, as well as one individual of *G. pallidiflora* that does not synthesize glycyrrhizin[29]. Hybridization is known to cause phylogenetic incongruence between different gene trees[30–32]. In our study, phylogenetic incongruence was indeed observed between concatenated and coalescent analyses based on single-copy nuclear genes (Supplementary Fig. S6), and between both nuclear phylogenies and maternally inherited plastome phylogeny (Fig. 2a), suggesting that hybridization occurred frequently between *Glycyrrhiza* species. Additionally, our species network inference (Fig. 2b), gene flow test (Fig. 3a), and population structure (Fig. 3b, c) all provided genetic evidence supporting the formation of the Ggu and Gug groups through hybridization between Gg and Gu (Fig. 5). These findings corroborate previous studies that reported frequent natural hybridization between Gg and Gu[33], provide genetic evidence that these hybrids with diverse genetic compositions mediate a genetic bridge between Gg and Gu, contributing to the maintenance of genetic diversity within *Glycyrrhiza* species (Fig. 5). Through our field investigations conducted in Central and East Asia, we confirmed that Gg is distributed in regions closer to Europe, such as Uzbekistan, while Gu has a distribution range closer to East Asia, including Mongolia and China. Furthermore, hybrid origin groups, Ggu and Gug, were widely distributed in Central Asia, including Russia, Kyrgyzstan, and Kazakhstan, which is consistent with previously reported findings[14]. In desert regions, where allogamy (outcrossing) pollinators are limited, Ggu and Gug have contributed to maintaining genetic and reproductive compatibility, thereby supporting the genetic diversity in the *Glycyrrhiza* species.

Legumes in the subfamily Papilionoideae have been reported to be predominantly autogamous or even cleistogamous due to their complicated flower morphology, as observed in genera such as *Macroptilium*, *Crotalaria*, and *Vigna*[34–38]. Although no studies have been conducted specifically on the reproductive system of *Glycyrrhiza* species, significantly low heterozygosity (Fig. 4a), high genetic differentiation (high F_{ST} in Fig. 4b), and complicated flower morphology (Supplementary Fig. S9) align with the other Papilionoideae species, implying that *Glycyrrhiza* species are likely to be autogamous in modern times, like other Papilionoideae species. However, the conditions during the Middle Pleistocene are estimated to have been different from those of modern times. According to changes in the historical effective population size of *Glycyrrhiza* species, notable fluctuations in population size, particularly in Gu and Ggu, have been inferred beginning around 2 Ma (Fig. 4c). This period coincides with the mid-Pleistocene climate transition (MPT; ~1.2 Ma to 0.4 Ma), characterized by prolonged glacial cycles lasting approximately 0.1 Ma, followed by asymmetric cycles with extensive cooling, the formation of thick ice sheets, and a rapid transition from extreme glacial conditions to warm interglacial periods[39,40]. The arid climate during the glacial periods likely provided favorable conditions for *Glycyrrhiza* species, which are adapted to steppe environments, allowing them to thrive. In contrast, during the same period, the reduction in effective population size for *Glycyrrhiza* collected from Central Asia (UZ01 and UZ02 of the Gg and KG01 of the Gug group) implies that these populations faced challenges due to glacial influences[14]. It has been discussed that the restriction of river flow effects in parts of Central Asia, such as Kazakhstan and the Altai mountains, may have led to more humid conditions, hindering the thriving of *Glycyrrhiza* species in that region[14].

In conclusion, our findings provide evidence that *Glycyrrhiza* species thrived during the Middle Pleistocene, and through active gene flow, various hybrids with diverse genetic compositions (Ggu1, Ggu2, Gug, and etc.) were formed and distributed across Central and East Asia. These hybrids are like genetic bridges between Gu and Gg, contributing to the maintenance of genetic diversity in these species. However, the current low heterozygosity and high genetic differentiation observed in Gu, Gg, and their hybrids imply that these *Glycyrrhiza* populations have been geographically, ecologically, or environmentally isolated for a long period, leading to a prolonged interruption of gene flow. We believe that this study enhances our understanding of the genetic characteristics of *Glycyrrhiza* species and contributes to our current knowledge of the genetic threats faced by *Glycyrrhiza* from a conservation perspective. Future studies should involve more individuals collected from a broader range of regions to examine the genetic diversity within *Glycyrrhiza* species. Additionally, high-quality genomes of other species, such as *G. glabra*, *G. inflata*, and *G. aspera*, as well as their hybrids, should be studied to further advance our understanding of these species.

Materials and methods

Sample collection and sequencing

We collected 29 individuals representing four *Glycyrrhiza* species—*G. uralensis*, *G. glabra*, *G. korshinskyi*, and *G. pallidiflora*—from seven countries in Central Asia (Supplementary Table S1). All materials were collected with permission in accordance with the regulations and guidelines of the respective countries and regions. The collected plants were identified by Dr. Jeonghoon Lee, and voucher specimens were deposited in the Korea Medicinal Resources Herbarium (KMRH) (Supplementary Table S1). Among these, one individual of *G. uralensis* (CN01) was selected for the reference genome assembly, and genome-wide variations were analyzed for all 29 individuals. Genomic DNA was extracted from fresh leaf tissues using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany), following the manufacturer's protocols. Whole-genome sequencing libraries were prepared using the TruSeq Nano DNA Sample Prep Kit (Illumina, CA, USA) with an insert size of 350 bp. Paired-end sequencing (2 × 150 bp) was performed on the NovaSeq6000 platform (Illumina, CA, USA). For the quality control of Illumina short-reads, we trimmed adaptors and low-quality reads using Trimmomatic ver. 0.38[41], with parameters as 'ILLUMINACLIP:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 HEADCROP:15 MINLEN:60'. Error correction and contaminant removal using tadpole.sh and bbsplit.sh scripts in BBtools ver. 38.26. After pre-processing, cleaned short-reads were used for downstream analysis. For the reference genome assembly, genomic DNA libraries were prepared using the ligation sequencing kit 1D (SQK-LSK109, Oxford Nanopore Technologies, UK) from the same leaf sample (CN01) used for WGS. We applied a single 1D flow cell on a PromethION sequencer, and base-calling for the PromethION platform was conducted using Minknow ver. 19.10.2 and Filp-Flop HAC model 29. Read quality was assessed using NanoPlot ver. 1.0.0, and adaptor sequences were removed with Porechop ver. 0.2.4. Hi-C data was produced with Hinf1 and DpnII as restriction enzymes.

Size Estimation and assembly of *Glycyrrhiza uralensis* genome

To efficiently identify variants across 29 *Glycyrrhiza* individuals and enhance our understanding of the *Glycyrrhiza* genome, we assembled the *G. uralensis* genome. The k-mer frequency analysis for genome size estimation was performed using Jellyfish ver. 2.2.4[42] and GenomeScope ver. 1.0.0[43] with trimmed Illumina sequences (CN01). The genome was assembled into contigs using Wtdbg2 ver. 2.5[43], followed by four rounds of polishing with Racon ver. 1.4.3 using the parameters '-m 8 -x -6 -g -8 -w 500'[44]. Additional polishing was performed with Medaka ver. 0.8.1 (https://github.com/nanoporetech/medaka) using a pre-trained Flip-Flop model. Two further rounds of short-read-based polishing were completed using Pilon ver. 1.22[45]. Hi-C sequencing data were integrated to order scaffolds into pseudo-molecules, utilizing the TRITEX pipeline[46], with the assembly published by Rai et al.[16] as the reference guide map. Genome completeness was evaluated using BUSCO analysis with Fabales_odb10 ver. 5.2.1[47].

Repeat and protein-coding gene annotation

The repeat library of the assembled genome was constructed using RepeatModeler ver. 2.0[48], and repeat annotation was performed with RepeatMasker ver. 4.1.0[49]. For gene annotation, RNA-sequence data from NCBI GenBank (accession numbers: DRR006519–DRR006526)[50] were utilized. The RNA-seq data were trimmed using Trimmomatic and de novo assembled with Trinity ver. 2.9.1[51]. To predict the gene models, trimmed RNA sequencing data were mapped onto the repeat-masked genome using HISAT2[52], and primary gene models were constructed with the BRAKER2 pipeline. These primary gene models, evidence from assembled transcripts, and the Swiss-Prot database[53] were integrated using the MAKER[54].

Evolutionary and comparative analyses of the *G. uralensis* genome

We downloaded legume genomes from the Pytozome database (https://phytozome-next.jgi.doe.gov/), including *Arachis duranensis* (V14167.gnm1), *Cicer arietinum* (492 v1.0), and *Lotus japonicus* (571_Lj1.0v1). The *Medicago truncatula* genome (ver. 4.0) was downloaded from NCBI GenBank (accession number: GCF000219495.3). Syntenic blocks were identified using the MCScanX algorithm[55]. Ks values were calculated using the "add_kaks_to_synteny.pl" script in the MCScanX package. Gene duplication events were identified with OrthoFinder ver. 2.5.4[56]. Functional annotation of *G. uralensis*-specific duplicated genes was conducted through homology searches against *Arabidopsis thaliana* genes. Gene enrichment analysis was performed using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) tool[57], applying enrichment criteria p-value < 0.01 and fold enrichment > 1.5.

Nuclear phylogenies and species network inference

To obtain nuclear single-copy genes, we used GFFRead[58] to generate fasta sequences corresponding to genes from the genome annotations of *G. uralensis* (CN01) in this study. Trimmed NGS files were used to assemble individual alignments with HybPiper[59]. HybPiper, developed under Biopython[60], utilized SPAdes[61] to initially execute read sorting via the BWA ver. 0.7.17[62] using the obtained genes. Following this preprocessing step, a *de novo* assembly was performed with SPAdes for each gene individually, utilizing the reads identified during the alignment step. The assembled gene sequences through HybPiper were used to gene clustering with OrthoFinder ver. 2.3.11[56]. Single-copy orthologous genes were then retained for further analysis. The selected single-copy orthologous genes were aligned using MAFFT[63] and concatenated into a single matrix. This concatenated dataset was used to construct a maximum likelihood (ML) tree using IQ-TREE ver. 2.3.6[64] with 1000 ultrafast bootstrap replicates and automatic selection of the best-fit substitution model (TEST). For each single-copy gene, an ML tree was also constructed using IQ-TREE with the same parameters used for the concatenated dataset. Using the maximum pseudo-likelihood method, we inferred species networks that model incomplete lineage sorting and gene flow (hybridization and introgression). The species network was inferred using PhyloNet ver. 3.8.0[65] with the command 'InferNetwork_MPL'. The species network inference was conducted with setting of 0 to 3 reticulations and optimizing branch lengths and inheritance probabilities (γ) under the pseudo-likelihood. The inferred species networks were displayed using Dendroscope ver. 3.6.3.

Plastid genome assembly and phylogeny

Using re-sequencing data of 29 *Glycyrrhiza* individuals, maternally inherited plastid genomes were assembled with the dnaLCW method[66,67]. Sequence trimming and *de novo* assembly of plastid genomes were performed using the CLC assembly cell ver. 4.21 (CLC Bio, Denmark). Trimmed reads were assembled with varying overlapping distances ranging from 150 to 500 bp. A previously reported plastid genome of *G. uralensis* (MN709826) was used as a reference. Contigs of plastid genomes were extracted using MUMmer[68] and BLASTZ[69], The assembled contigs were manually curated and further assembled into a single draft sequence. Gene annotation of the assembled plastid genomes was performed using GeSeq[70], followed by manual correction using Artemis[71]. Circular map of plastid genomes was drawn using OGDRAW[72]. Plastid genome sequences of 29 *Glycyrrhiza* samples were used for phylogenetic reconstruction. *G. pallidiflora* (KR05) was selected as an outgroup. All plastid genome sequences were aligned using MAFFT. Phylogenetic analyses were performed using IQ-TREE with 1000 ultrafast bootstrap replicates and the automatic selection of the best-fit substitution model. The results of phylogenetic analyses were visualized using iTOL[73].

Variation calling and population analysis

Structural variation calling was performed by aligning two *G. uralensis* genomes using minimap2 ver. 2.24[74] and identifying variants with SYRI. The aligned sequences were visualized as a dot plot with D-genies[75]. For SNV calling, we aligned the cleaned NGS reads with the *G. uralensis* (CN01) genome using BWA-MEM ver. 0.7.17[76] and processed with Novosort ver. 3.06.05 (http://www.novocraft.com). SNVs were called using GATK ver. 3.8[77], and filtered out those with Q scores of less than 30 using SAMtools ver. 1.9[78]. SNV annotation was conducted using SnpEff ver. 4.3[79]. To reduce false-positives from SNV and genotype calling, we conducted numerous filtering steps: (1) removal of SNVs showing extremely low (< 100 reads across all samples per species) or high (> 1,200 reads across all samples per species) read depths; (2) removal of SNVs in repetitive regions; (3) removal of SNVs with more than two alleles; (4) removal of SNVs at or within 5 bp from any indel; (5) use of at least three genotypes with a quality score (GQ) > 10; and vi) removal of SNVs showing significant deviation from Hardy-Weinberg equilibrium ($p < 0.001$). Principal component analysis (PCA) was performed using PLINK ver. 1.9[80] with default parameters, and the resulting PCA plot was generated with ggplot2 in the R package. Population structure analysis was conducted using ADMIXTURE ver. 1.3.0[81], with the optimal number of subpopulation groups determined based on the lowest cross-validation error, with k ranging from two to six.

Gene flow, heterozygosity, historical dynamic analyses

HyDe[82] was used to assess the gene flow, including hybridization and introgression, among 29 *Glycyrrhiza* samples. HyDe is similar to the ABBA-BABA test (*D*-statistic) and has been shown to provide helpful information for detecting hybridization among three ingroups based on phylogenetic invariants[82]. To conduct the HyDe analysis, we used SNV data and selected *G. pallidiflora* (KR05) as an outgroup. Based on phylogenetic results of nuclear single-copy genes and plastid genomes, the ingroups were defined as follows: *G. glabra* (Gg) clade, *G. uralensis* (Gu) clade, two *G. glabra* × *uralensis* (Ggu 1 and 2) clades, and *G. uralensis* × *glabra* (Gug) clade. If gene flow has occurred between two of ingroups, the P-value is statistically significant (< 0.05), and the inheritance probability (γ) is estimated to range between 0 and 1, indicating varying degrees of hybridization between the two putative parents (γ from P1 and 1-γ from P2). To visually demonstrate the statistical substantiation of detected gene flow, violin plots were drawn to highlight the variance and central tendency of the simulation results. The heterozygosity of 29 *Glycyrrhiza* individuals was calculated with ANGSD ver. 0.9.3[83]. Whole-genome SNVs were used to estimate the demographic history of *Glycyrrhiza* species by applying PSMC analysis[84]. PSMC ver. 0.6.5 analysis was conducted using 100 bootstrapping rounds with 0.6×10^{-8} substitutions per site per generation over a generation time of three years, based on previous studies indicating that *Glycyrrhiza* species require at least three years to reach maturity and accumulate key secondary metabolites in their roots[16,85,86].

Sample collection and experiment statement

All the plant materials in this study were collected with permission in accordance with the regulations and guidelines of the respective countries and regions, and all the methods in this study comply with relevant institutional, national, and international guidelines and legislation.

Data availability

All sequences generated in this study, including PromethION long reads and Illumina short reads, were deposited in the NCBI SRA under BioProject PRJNA787406.

Received: 24 January 2025; Accepted: 25 February 2025
Published online: 13 March 2025

References

1. Chen, M. et al. Exploration in the mechanism of action of licorice by network Pharmacology. *Molecules* **24**, 2959 (2019).
2. Tian, M., Yan, H. & Row, K. H. Extraction of glycyrrhizic acid and glabridin from licorice. *Int. J. Mol. Sci.* **9**, 571–577 (2008).
3. Gupta, S., Sharma, R., Pandotra, P., Jaglan, S. & Gupta, A. P. Chromolithic method development, validation and system suitability analysis of ultra-sound assisted extraction of glycyrrhizic acid and glycyrrhetinic acid from *Glycyrrhiza glabra*. *Nat. Prod. Commun.* **7**, 1934578X1200700808 (2012).
4. Guo, J., Yang, C., Yang, J. & Yao, Y. Glycyrrhizic acid ameliorates cognitive impairment in a rat model of vascular dementia associated with oxidative damage and Inhibition of voltage-gated sodium channels. *CNS Neurol. Disord -Drug Targ.* **15**, 1001–1008 (2016).
5. Han, S., Sun, L., He, F. & Che, H. Anti-allergic activity of glycyrrhizic acid on IgE-mediated allergic reaction by regulation of allergy-related immune cells. *Sci. Rep.* **7**, 7222 (2017).
6. Feldman, M., Santos, J. & Grenier, D. Comparative evaluation of two structurally related flavonoids, Isoliquiritigenin and Liquiritigenin, for their oral infection therapeutic potential. *J. Nat. Prod.* **74**, 1862–1867 (2011).
7. Gaur, R., Yadav, K. S., Verma, R. K., Yadav, N. P. & Bhakuni, R. S. In vivo anti-diabetic activity of derivatives of Isoliquiritigenin and Liquiritigenin. *Phytomedicine* **21**, 415–422 (2014).
8. Zhai, K. F. et al. Liquiritin from *Glycyrrhiza uralensis* attenuating rheumatoid arthritis via reducing inflammation, suppressing angiogenesis, and inhibiting MAPK signaling pathway. *J. Agric. Food Chem.* **67**, 2856–2864 (2019).
9. Bao, B. J., Larsen, K. & Glycyrrhiza Missouri Botanical Garden Press. In *Flora of China* Vol. 10 (eds Wu, Z. Y., Hong, D. Y., & Raven, P. H.) 509–511 (2010).
10. Hayashi, H. et al. Field survey of *Glycyrrhiza* plants in central Asia (2). Characterization of phenolics and their variation in the leaves of *Glycyrrhiza* plants collected in Kazakhstan. *Chem. Pharm. Bull.* **51**, 1147–1152 (2003).
11. Duan, L. et al. Species delimitation of the liquorice tribe (Leguminosae: Glycyrrhizeae) based on phylogenomic and machine learning analyses. *J. Syst. Evol.* **61**, 22–41 (2023).
12. Duan, L. et al. Chloroplast phylogenomics reveals the intercontinental biogeographic history of the liquorice genus (Leguminosae: *Glycyrrhiza*). *Front. Plant. Sci.* **11**, 793 (2020).
13. Grigorev, Y. S. Towards systematics of the genus *Glycyrrhiza* L. *Izvestija Glavnogo Botaničeskogo Sada SSSR*. **29**, 92–98 (1930).
14. Hantemirova, E. V., Belyaev, A. Y., Korchagina, O. S. & Laletina, I. A. Genetic differentiation and phylogenetic relationships of *Glycyrrhiza glabra* L., *G. uralensis* Fisch., and *G. korshinskyi* Grig. inferred from chloroplast DNA variation. *Russ. J. Genet.* **56**, 810–821 (2020).
15. Mochida, K. et al. Draft genome assembly and annotation of *Glycyrrhiza uralensis*, a medicinal legume. *Plant. J.* **89**, 181–194 (2017).
16. Rai, A. et al. Chromosome-scale genome assembly of *Glycyrrhiza uralensis* revealed metabolic gene cluster centred specialized metabolites biosynthesis. *DNA Res.* **29**, dsac043 (2022).
17. Zhang, J. T., Xu, B. & Li, M. Genetic diversity of populations of an endangered medicinal plant species (*Glycyrrhiza uralensis*) in different environments of North China. *J. Med. Plants Res.* **4**, 830–836 (2010).
18. Um, Y. et al. Genetic diversity analysis of *Glycyrrhiza uralensis* using 8 novel polymorphic microsatellite markers. *J. Plant. Biotechnol.* **43**, 174–180 (2016).
19. Esmaeili, H., Karami, A., Hadian, J., Ebrahimi, S. N. & Otto, L. G. Genetic structure and variation in Iranian licorice (*Glycyrrhiza glabra* L.) populations based on morphological, phytochemical and simple sequence repeats markers. *Ind. Crop Prod.* **145**, 112140 (2020).
20. Liu, Y. et al. Genetic structure and diversity of *Glycyrrhiza* populations based on transcriptome SSR markers. *Plant. Mol. Biol. Rep.* **37**, 401–412 (2019).
21. Binder, B. M. Ethylene signaling in plants. *J. Biol. Chem.* **295**, 7710–7725 (2020).
22. Yan, B. et al. A review on the plant resources of important medicinal licorice. *J. Ethnopharmacol.* **301**, 115823 (2023).
23. Lee, J. H. et al. Wongam', a licorice interspecific hybrid cultivar with high yield. *Korean J. Breed. Sci.* **52**, 454–459 (2020).
24. Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J. & Edwards, D. Plant pan-genomes are the new reference. *Nat. Plants* **6**, 914–920 (2020).
25. Khan, A. W. et al. Super-pangenome by integrating the wild side of a species for accelerated crop improvement. *Trends Plant. Sci.* **25**, 148–158 (2020).
26. Jayakodi, M., Schreiber, M., Stein, N. & Mascher, M. Building pan-genome infrastructures for crop plants and their use in association genetics. *DNA Res.* **28**, dsaa030 (2021).
27. He, Q. et al. A graph-based genome and pan-genome variation of the model plant *Setaria*. *Nat. Genet.* **55**, 1232–1242 (2023).
28. Shi, J., Tian, Z., Lai, J. & Huang, X. Plant pan-genomics and its applications. *Mol. Plant.* **16**, 168–186 (2023).
29. Kajiyama, K. et al. Flavonoids and isoflavonoids of chemotaxonomic significance from *Glycyrrhiza pallidiflora* (Leguminosae). *Biochem. Syst. Ecol.* **21**, 785–793 (1993).
30. Maddison, W. P. Gene trees in species trees. *Syst. Biol.* **46**, 523–536 (1997).
31. Folk, R. A., Soltis, P. S., Soltis, D. E. & Guralnick, R. New prospects in the detection and comparative analysis of hybridization in the tree of life. *Am. J. Bot.* **105**, 364–375 (2018).
32. Stull, G. W., Pham, K. K., Soltis, P. S. & Soltis, D. E. Deep reticulation: the long legacy of hybridization in vascular plant evolution. *Plant. J.* **114**, 743–766 (2023).
33. Zimnitskaya, S. A. State of the reproductive system of populations of species of the genus *Glycyrrhiza* L.(Fabaceae). *Contemp. Probl. Ecol.* **2**, 392–395 (2009).
34. Gopinathan, M. C. & Babu, C. R. Breeding systems and pollination in *Vigna minima* (Leguminosae, Papilionoideae). *Plant. Syst. Evol.* **156**, 117–126 (1987).
35. Drewes, S. I. & Hoc, P. S. Morfología y desarrollo de Flores cleistógamas En *Macroptilium fraternum* (Fabaceae). *Kurtziana* **28**, 229–238 (2000).

36. Etcheverry, A. V., Alemán, M. M. & Fleming, T. F. Flower morphology, pollination biology and mating system of the complex flower of *Vigna caracalla* (Fabaceae: Papilionoideae). *Ann. Bot.* **102**, 305–316 (2008).
37. Etcheverry, A. V., Perez de Bianchi, S. M. & de Martin, D. Reproductive biology in the amphicarpic legume *Macroptilium panduratum* (Fabaceae, Papilionoideae). *Beiträge Zur Biol. Der Pflanzen* **72**, 181–197 (2001).
38. Etcheverry, A. V., Protomastro, J. J. & Westerkamp, C. Delayed autonomous self-pollination in the colonizer *Crotalaria micans* (Fabaceae: Papilionoideae): structural and functional aspect. *Plant. Syst. Evol.* **239**, 15–28 (2003).
39. Diester-Haass, L., Billups, K. & Lear, C. Productivity changes across the mid-Pleistocene climate transition. *Earth Sci. Rev.* **179**, 372–391 (2018).
40. Brovkin, V., Calov, R., Ganopolski, A. & Willeit, M. Mid-Pleistocene transition in glacial cycles explained by declining CO_2 and regolith removal. *Sci. Adv.* **5**, eaav7337 (2019).
41. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
42. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
43. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* **17**, 155–158 (2020).
44. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de Novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
45. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One* **9**, e112963 (2014).
46. Marone, M. P., Singh, H. C., Pozniak, C. J. & Mascher, M. A technical guide to TRITEX, a computational pipeline for chromosome-scale sequence assembly of plant genomes. *Plant. Methods* **18**, 128 (2022).
47. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
48. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2020).
49. Tarailo-Graovac, M. & Chen, N. Using repeatmasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform.* **25**, 4–10 (2009).
50. Ramilowski, J. A. et al. *Glycyrrhiza uralensis* transcriptome landscape and study of phytochemicals. *Plant. Cell. Physiol.* **54**, 697–710 (2013).
51. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
52. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
53. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement trembl in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
54. Cantarel, B. L. et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).
55. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49–e49 (2012).
56. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
57. Sherman, B. T. et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res.* **50**, W216–W221 (2022).
58. Pertea, G. & Pertea, M. GFF utilities: GffRead and GffCompare. *F1000Research* 9 (2020).
59. Johnson, M. G. et al. HybPiper: extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Appl. Plant. Sci.* **4**, 1600016 (2016).
60. Cock, P. J. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422 (2009).
61. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
62. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
63. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
64. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
65. Than, C., Ruths, D. & Nakhleh, L. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinform.* **9**, 322 (2008).
66. Kim, K. et al. Comprehensive survey of genetic diversity in Chloroplast genomes and 45S NrDNAs within *Panax ginseng* species. *PLoS One* **10**, e0117159 (2015).
67. Kim, K. et al. Complete Chloroplast and ribosomal sequences for 30 accessions elucidate evolution of *Oryza* AA genome species. *Sci. Rep.* **5**, 15655 (2015).
68. Kurtz, S. et al. Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
69. Schwartz, S. et al. Human–mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003).
70. Tillich, M. et al. GeSeq–versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* **45**, W6–W11 (2017).
71. Carver, T., Harris, S. R., Berriman, M., Parkhill, J. & McQuillan, J. A. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* **28**, 464–469 (2012).
72. Greiner, S., Lehwark, P. & Bock, R. OrganellarGenomeDRAW (OGDRAW) version 1.3. 1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* **47**, W59–W64 (2019).
73. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
74. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
75. Cabanettes, F. & Klopp, C. D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **6**, e4958 (2018).
76. Li, H. Aligning sequence Reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv* 13033997 (2013).
77. McKenna, A. et al. The genome analysis toolkit: A mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
78. Li, H. et al. The sequence alignment/map format and samtools. *Bioinformatics* **25**, 2078–2079 (2009).
79. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
80. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
81. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based Estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).

82. Blischak, P. D., Chifman, J., Wolfe, A. D. & Kubatko, L. S. HyDe: a python package for genome-scale hybridization detection. *Syst. Biol.* **67**, 821–829 (2018).
83. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC Bioinform.* **15**, 1–13 (2014).
84. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
85. Chiyo, N. et al. Glycyrrhizin production in licorice hairy roots based on metabolic Redirection of triterpenoid biosynthetic pathway by genome editing. *Plant. Cell. Physiol.* **65**, 185–198 (2023).
86. Kozhuharova, A. et al. Establishment of an ex situ collection of *Glycyrrhiza glabra* L. as a prerequisite for field cultivation in Bulgaria. *Horticulturae* **10**, 86 (2024).

Author contributions

J.K., J.L., Y.S., J.B., and T.J.Y. planned and designed the research; J.L., Y.K., Y.L., and Z.Z collected plant materials. J.S.K., J.K., H.S., D.K., S.H.L., J.C., H.K., and M.K. analyzed the data. J.S.K., J.K., J.L., Y.S., J.B., and T.J.Y. wrote the manuscript.

Funding

This work was supported by the "Bio-valley for Academy Industry Convergence" in Osong and "Cooperative Research Program for Agriculture Science & Technology Development (Grant numbers: PJ01424601, PJ017134), funded by Rural Development Administration, Republic of Korea.

Declarations

Competing interests

J.B. is the chief executive officer of Clinomics Inc., and all other authors declare that they have no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-92115-4.

Correspondence and requests for materials should be addressed to Y.K.S., J.B. or T.-J.Y.

Reprints and permissions information is available at www.nature.com/reprints.

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