An *et al. BMC Plant Biology* (2026) 26:686
https://doi.org/10.1186/s12870-026-08422-z

BMC Plant Biology

RESEARCH Open Access

Plastome evolution and phylogenomics of *Glycine* (Leguminosae: Papilionoideae)

Kai-Lun An[1,2], Wei Gu[1,2], Qing Lu[1,2], Ding-Jie Wang[1,2], Zuo-Ying Xiahou[1,2], Xiao-Gang Fu[1,2], Libo Zhao[1,2], Zhi-Yun Yang[1,2], Robert Henry[3,4], Jian Wang[3,5*], Rong Zhang[1,2*] and Ting-Shuang Yi[1,2*]

Abstract

The genus *Glycine* Willd. is economically important due to the soybean, a major global crop for oil and protein. However, comprehensive phylogenetic studies of this genus are lacking, and plastome evolution remains poorly understood. In this study, we conducted comparative plastomic and phylogenetic analyses using 42 complete *Glycine* plastomes. This dataset encompassed 28 newly sequenced plastomes, two reassembled from public raw data, and 12 obtained from GenBank, collectively representing 15 of the 29 species in this genus. Comparative analyses revealed these *Glycine* plastomes range from 152,225 to 152,972 bp in length, and displayed highly conserved quadripartite structure, gene content (111 genes), and collinearity. Nine protein-coding genes and six non-coding regions showed high sequence variability, were thus identified as promising markers for phylogenetics and species delimitation. Selection pressure analyses detected branch-specific positive selection ($d_N/d_S > 1$) in *rpo*B, *acc*D, and *rbc*L, supporting their potential roles in environmental adaptation. Phylogenomic analyses strongly supported the monophyly of *Glycine* and its two subgenera (*Soja* and *Glycine*), resolving three major clades within subgenus *Glycine*. However, several species were resolved as non-monophyletic, and the plastomic phylogeny conflicted with nuclear data, indicating cytonuclear discordance likely caused by hybridization, plastid capture, and/or incomplete lineage sorting. These findings highlight the complex evolutionary history of the genus. This study provides the most comprehensive plastome dataset for *Glycine* to date and offers new insights into its plastome evolution and cytonuclear discordance.

*Correspondence:
Jian Wang
jian.wang@detsi.qld.gov.au
Rong Zhang
zhangronga@mail.kib.ac.cn
Ting-Shuang Yi
tingshuangyi@mail.kib.ac.cn
[1]The Germplasm Bank of Wild Species & Yunnan Key Laboratory of Crop Wild Relatives Omics, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, St Lucia, QLD 4072, Australia
[4]VinUni Big Data Research Institute, VinUniversity, Hanoi, Vietnam
[5]Queensland Herbarium & Biodiversity Science, Brisbane, QLD 4066, Australia

Introduction

The genus *Glycine* Willd. (Leguminosae: Papilionoideae) comprises two subgenera, *Soja* and *Glycine*, with a total of 29 herbaceous species. Subgenus *Soja* contains two annual species: the wild soybean *G. soja* Siebold & Zucc., native to East Asia, and the cultivated soybean *G. max* (L.) Merr., which was domesticated from *G. soja* in China. Subgenus *Glycine* includes 27 perennial species, of which 23 are endemic to Australia, two (*G. tabacina* (Labill.) Benth. and *G. tomentella* Hayata) are distributed in southestern China, Japan, the islands of Southeast Asia, and Australia, one (*Glycine dolichocarpa* Tateishi & H.Ohashi) in Taiwan, China, and one (*G. koidzumii* Ohwi) in Japan. *Glycine*

species occupy a wide range of habitats, including deserts, sandy beaches, various forest types spanning temperate to subtropical zones, and moist shrublands, demonstrating exceptional environmental plasticity [1, 2]. They exhibit enhanced tolerance to abiotic stresses (e.g., drought, salinity) and robust resistance against biotic stresses (e.g., pathogens, herbivores), making them valuable genetic resources for soybean improvement [3–5].

The genus *Glycine* Willd. was first established by Linnaeus with *G. javanica* L. as the type species [6], a designation retained by Willdenow [7]. De Candolle [8] controversially transferred many Phaseoleae species into *Glycine* L., substantially expanding *Glycine* to 286 species. Hermann [9] subsequently revised *Glycine*, retaining 10 species and dividing the genus into three subgenera: subg. *Soja*, subg. *Glycine*, and subg. *Leptocyamus*. Later, Verdcourt [10] demonstrated the original type of *G. javanica* belonged to *Pueraria* DC and designated *G. clandestina* J.C.Wendl. as the new type species [10]. Verdcourt [10] and Lackey [11] revised the genus into two subgenera: subg. *Soja* and subg. *Glycine*, a classification widely adopted in subsequent studies.

A robust phylogenetic framework is essential for elucidating evolutionary processes such as speciation, diversification, coevolution, and biogeographic patterns [12]. Despite the prevalent application of nuclear genomic data in phylogenetics, plastomes retain indispensable for phylogenetics due to their highly conserved structure and gene content, uniparental inheritance and low recombination rates, which reduce paralog interference and make them valuable for resolving complex reticulate evolutionary histories [13–20]. Comparisons between plastome and nuclear genome data could reveal evolutionary discordance caused by processes like hybridization [21–24], which are poorly studied in *Glycine*. Although most legume plastomes have the typical quadripartite structure, significant structural variations exist in some lineages, such as extensive rearrangements and IR loss in the IR-lacking clade (IRLC) [25]. *Glycine* plastomes are structurally conserved based on limited samples, allowing us to investigate structural variation and sequence-level evolution with broader sampling.

Previous phylogenetic studies using a few plastid loci robustly support the division of *Glycine* into two subgenera but offer limited phylogenetic resolution [26–29]. Subsequent plastid and nuclear phylogenomic analyses significantly improved resolution but suffered from limited species sampling [30–33]. Recent nuclear studies with broader sampling, resolved major clades and interspecific relationships of the genus [5, 31]. Nevertheless, these nuclear phylogenies exhibit topological conflicts with plastome-based phylogenies. For instance, nuclear data consistently position *G. falcata* Benth. as a sister to all other species within the subgenus *Glycine*, whereas plastid data place it within a clade of this subgenus [5, 30, 31]. This widespread cytonuclear discordance likely results from hybridization, introgression, and/or incomplete lineage sorting (ILS). Furthermore, previous nuclear studies have revealed multiple non-monophyletic *G. tomentella* and *G. tabacina* [27, 31], yet comparable plastome-level studies are still lacking. Therefore, it is crucial to expand species- and individual-level sampling for plastome-based phylogenomic studies.

Plastids serve as the primary organelles for anabolic processes in plant cells [34]. Most plant plastomes possess a conserved quadripartite structure, comprising a large single-copy (LSC) region, a small single-copy (SSC) region, and two inverted repeat (IR_A and IR_B) regions. Plastome size usually ranges from 115 to 165 kb in length and contains 101 to 118 genes primarily involved in photosynthesis, transcription, and translation [35, 36]. Plastome evolution mirrors the adaptive history of plants, with gene loss, structural rearrangement, and horizontal transfer revealing adaptations to diverse ecological environments [37, 38]. Additionally, IR regions play a vital role in maintaining plastome stability, and their expansion or contraction provides important phylogenetic signals [39]. A notable case of such an evolutionary signal is the IRLC clade within the subfamily Papilionoideae, characterized by the complete loss of the IR_A region, which renders it evolutionarily and genetically unique [25]. Furthermore, most Papilionoideae, except for a few early diverging lineages, exhibit a 50-kb inversion in the LSC region [40–42]. Doyle et al. [28] demonstrated that the genus *Glycine* shares this 50-kb inversion with most Papilionoideae species. Sherman et al. [30] analyzed plastome structure in seven *Glycine* species, revealing relatively conserved IR boundaries and no structural rearrangements in perennial species. However, comprehensive studies on structural variations and evolutionary patterns across a broad sampling of *Glycine* plastomes are still needed.

In this study, we newly sequenced and assembled plastomes from 28 accessions of 11 species representing both *Glycine* subgenera and all three recognized clades within the subgenus *Glycine*. By integrating these newly sequenced with published *Glycine* plastomes, we analyzed their plastome structure and sequence divergence patterns, and reconstructed a robust plastome-based phylogenomic framework. Our study aims to (1) characterize structural variations and nucleotide divergence patterns across *Glycine* plastomes and (2) elucidate maternal lineage evolutionary trajectories within this genus.

Materials and methods

Taxa sampling

To ensure broad taxonomic and geographic representation, a total of 42 samples representing 15 *Glycine* species from the major lineages of the genus were strategically selected, including 28 newly sequenced accessions representing 11species (detailed information provided in the Supplemental Table S1). For the newly sequenced samples, leaf materials were obtained from herbarium collections at the Chinese National Herbarium, Institute of Botany, Chinese Academy of Sciences (PE, China) and the Queensland Herbarium, Queensland Herbarium and Biodiversity Science (BRI, Australia) under standard sampling protocols (six from PE, and remainder from BRI; Supplemental Table S1). The plastomes of *G. clandestina*, *G. curvata*, *G. latifolia*, *G. latrobeana*, and *G. microphylla* are firstly reported. Voucher details are in Supplemental Table S1; the collection details and GenBank accession numbers for sequences from NCBI are provided in Supplemental Table S2.

DNA extraction and sequencing

For the 28 newly sequenced plastomes, total DNA was extracted from ~ 15–20 mg (equivalent to ~ 1 cm^2) silica-gel-dried or specimen leaf tissue using the CTAB method [43]. DNA was quantified using a Qubit Fluorometer. Genomic DNA libraries were constructed using the NEBNext® Ultra™ II DNA Library Prep Kit (New England Biolabs, USA) following the manufacturer's instructions. Library quality was assessed by measuring concentration using the Qubit fluorometer, with a minimum threshold of 10 ng/μL required for sequencing (libraries ≥ 20 ng/μL were considered high-quality). The size distribution of the library fragments was assessed using an Agilent 2100 Bioanalyzer to ensure that fragments were predominantly concentrated within the 200–500 bp range. Libraries were purified and size-selected using AMPure XP magnetic beads, followed by two washes with 80% ethanol to remove impurities. The libraries were subsequently sequenced on the Illumina HiSeq 4000 platform with 150-bp paired-end reads at the Molecular Biology Experiment Platform, Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences (KIB). Raw reads were quality-checked with FastQC and filtered to remove adapter sequences, reads with > 10% unknown (N) bases, and reads where > 50% of bases had a Phred quality score (Q) ≤ 20. The resulting high-quality clean data (with average Q30 scores > 95%) were used for subsequent de novo assembly.

Plastome assembly and annotation

Plastome assembly was conducted using GetOrganelle v1.7.7 [44] with optimized parameters (-k 21,45,65,85,105,127; -R 15). Two plastomes of *G. remota* (SRR18315596) and *G. max* × *G. soja* (SRR9714351) were reassembled from public raw data. Assembled graphs were visualized and confirmed using Bandage v0.8.1 [45]. A successful assembly was defined by a single closed circular path and the conserved quadripartite plastome structure. Genome annotation was performed using the Plastid Genome Annotator (PGA) with default parameters [46], with *G. max* (NC_007942) from NCBI as the reference. Manual refinement of annotations was carried out in Geneious v9.1.8 [47] to ensure precise demarcation of start and stop codons and confirm comprehensive gene coverage. Physical maps of the 28 newly sequenced plastomes were generated using the online tool OGDRAW [48].

Genome comparison and sequence divergence analysis

Whole-plastome collinearity was analyzed using the progressiveMauve plugin in Geneious v9.1.8 [47] with default parameters. Boundaries of the LSC, SSC, and IR regions were visualized with the online web tool CPJSDRAW [49]. Whole-plastome alignment and visualization were performed using mVISTA [50] under the shuffle-LAGAN alignment algorithm with *G. max* (NC_007942) as reference. Nucleotide diversity analysis of protein-coding, intron and intergenic regions (> 200 bp) was calculated using DNASP v6.12.03 [51]. These regions were extracted using the python script "get_annotated_regions_from_gb.py" (https://github.com/Kinggerm/PersonalUtilities).

Repeat sequences identification and codon usage analysis

Simple sequence repeats (SSRs) were identified using MISA v2.1 [52, 53] with thresholds of 10, 5, 4, 3, 3, and 3 for mono-, di-, tri-, tetra-, penta-, and hexanucleotides, respectively. Long repeats (forward, reverse, complement and palindromic; ≥ 30 bp, Hamming distance = 3) were detected with REPuter [54, 55]. One IR copy was removed from both analyses to avoid redundancy. Codon usage bias for all protein-coding genes (PCGs) across 42 *Glycine* plastomes was assessed by calculating codon counts and Relative Synonymous Codon Usage (RSCU) using CodonW v1.4.2 (http://codonw.sourceforge.net). RSCU reflects codon preference bias: values > 1.0 indicate overrepresented codons, < 1.0 underrepresented codons, and 1.0 denotes neutral usage (no bias). To further visually illustrate codon usage patterns, this study employed the online tool RSCU-Plot [56–59] (https://pcg-lab.shinyapps.io/RSCU-Plot/) to generate bar charts of RSCU values for individual samples, clearly presenting the distribution characteristics of overrepresented and underrepresented codons.

Selective pressure analysis

Protein-coding genes (PCGs) were aligned using the codon mode of the MAFFT plugin in PhyloSuite v1.2.3

[60], and converted to the "PML" format via the Convert Sequence Format tool. The 42 *Glycine* plastome sequences were aligned with MAFFT under default parameters, and a maximum likelihood (ML) tree was reconstructed with RAxML v8.2.12 [61] (with optimized parameters: -m GTRGAMMA -N 1000 -f a -x 12345 -p 12345).

With the ML tree and PML-formatted sequence files as inputs, the nonsynonymous (d_N) and synonymous (d_S) substitution rates of each plastid gene was assessed uing the CodeML program in PAML v4.10.7 [62]. For each PCG, the one-ratio model (assuming uniform ω across all branches; parameters: model = 0, NSites = 0) and two-ratio model (distinct ω for foreground vs. background branches; parameters: model = 2, NSites = 0) were employed to detect selection differences. Model fit was determined by Likelihood Ratio Tests (LRT). Eight foreground branches (Fig. 1) were labeled using EasyCodeML v1.0 [63], with all configurations applied per PAML manual specifications to all target branches.

Phylogeny analysis

Following previously constructed phylogenetic frameworks [64, 65], *Neonotonia wightii* (Wight & Arn.) J.A.Lackey and *Phylacium bracteosum* Benn. (Phaseoleae, Glycininae) were selected as outgroups for phylogenetic reconstruction. The 77 protein-coding regions (CDS) and 132 non-coding sequences (NCR) (excluding 1-bp *ndh*A-*ndh*H intergenic region) from all 42 *Glycine* plastomes were extracted using the python script

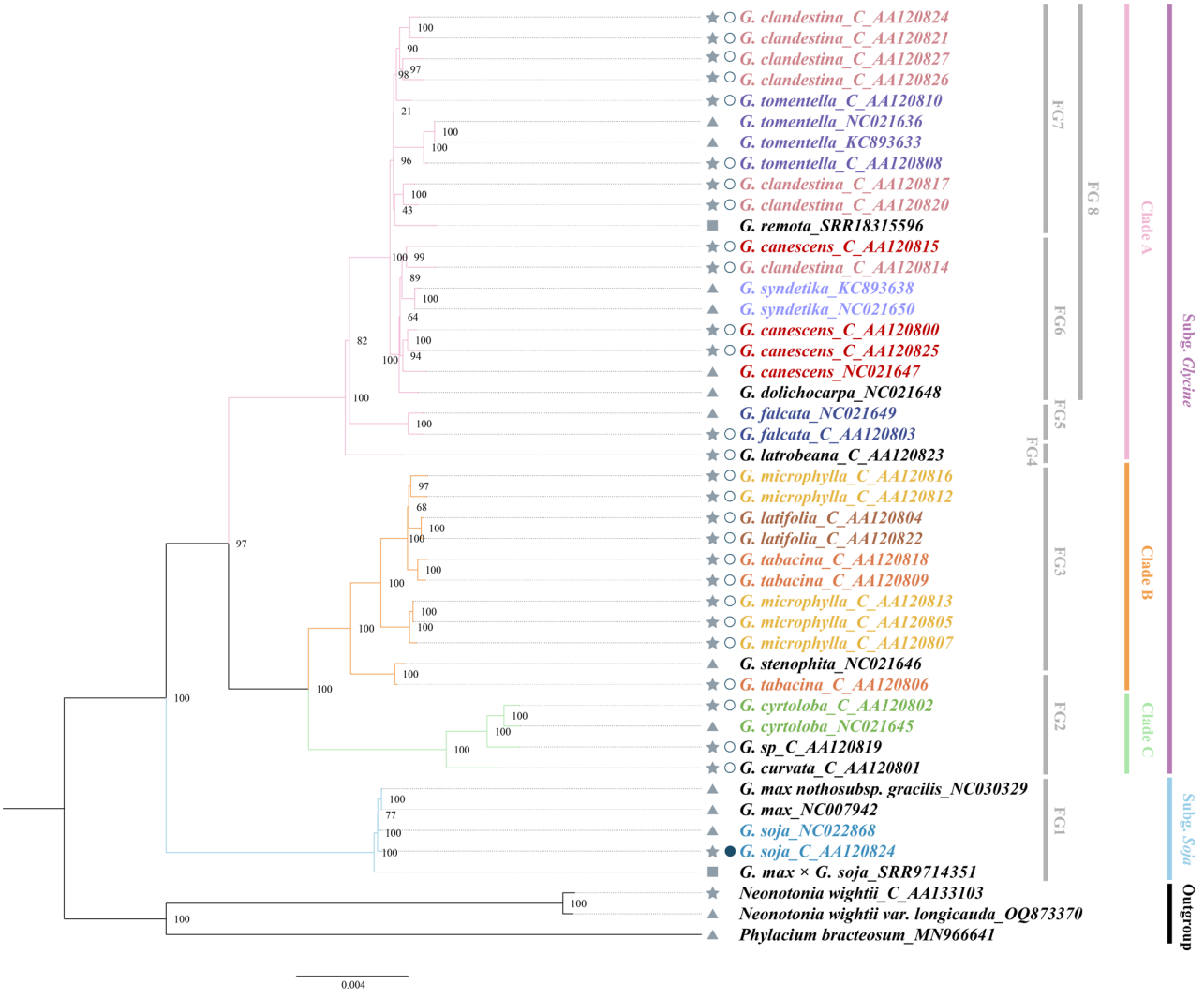

Fig. 1 Maximum Likelihood (ML) phylogenetic tree of *Glycine* based on the CP matrix. Taxa labels of the same color indicate different individuals of the same species. Two sets of symbols preceding the labels denote data sources and sampling locations, respectively. The first column presents data sources: stars, newly sequenced samples; squares, samples reassembled from published data; triangles, plastome sequences retrieved from NCBI. The second column presents sampling locations: open circles, Australia; solid circles, China; no symbols, published data with unavailable collection sites. Gray bars adjacent to taxa labels denote the specific foreground branches designated for the branch-model of the selection pressure analysis

"get_annotated_regions_from_gb.py" (https://github.com/Kinggerm/PersonalUtilities). Sequences were aligned using the MAFFT plugin in PhyloSuite v1.2.3 [60] with default parameters. Poorly aligned sites, defined as alignment columns with a high proportion of gaps or low sequence similarity, were automatically identified and removed using trimAl v1.4 under default parameters [66]. The resulting alignments were then concatenated in PhyloSuite to generate CDS, NCR, and complete plastome (CP, constructed by concatenating CDS and NCR sequences back-to-back, equivalent to a complete plastome with one IR region removed) datasets. The best nucleotide substitution model was selected with ModelFinder. The ML tree was reconstructed for each matrix using RAxML v8.2.12 [61], applying the GTRGAMMA model (recommended by RAxML) and GTR + F+I+G4 model (selected by ModelFinder) nucleotide substitution model, respectively. Bootstrap support (BS) was assessed with 1000 rapid replicates.

Results

Plastome features

All 28 newly sequenced *Glycine* plastomes exhibit a typical quadripartite structure, comprising LSC, SSC, and two IR regions (Fig. 2). These plastomes ranges in size from 152,225 to 152,972 bp (e.g., Fig. 3; LSC: 83,187–84,072 bp; SSC: 17,756–17,890 bp; IR: 25,405–25,575 bp), with GC content of 35.3–35.4% (Table S3). All 42 plastomes (28 newly sequenced and 14 published) show high collinearity, with identical gene content and arrangement (Fig. S1). These plastomes comprise 111 genes, including 77 protein-coding, 30 tRNA, and four rRNA genes (Fig. 2; Table 1, S3). Among these genes, 6 tRNA and 9 protein-coding genes contain a single intron, while *clp*P, *rps*12 and *ycf*3 contain two introns (Table 1).

Phylogeny inferences

Phylogenetic trees constructed from CDS, NCR, and CP matrices under two distinct models exhibited congruent topologies, with most nodes strongly supported (BS > 90%) (Fig. 1, S2, S3, S6). As the CP-based tree had the highest overall support, subsequent analyses were based on this topology. Monophyly of *Glycine* and its two subgenera (*Glycine* and *Soja*) were strongly supported (BS = 100%), as were clades A, B, and C within subgenus *Glycine* (BS = 100%). Within subgenus *Glycine*, Clades B and C formed a sister group (BS = 100%) and together were sister to Clade A (BS = 97%) (Fig. 1). Six species with multiple accessions were non-monophyletic: *G. clandestina*, *G. tomentella*, *G. canescens*, *G. microphylla*, and *G. tabacina* (subgenus *Glycine*), as well as *G. soja* (subgenus *Soja*). For the five non-monophyletic species in subgenus *Glycine* where geographic data was available, individuals from the same region did not form exclusive clades, suggesting the observed polyphyly is not merely a geographic artifact.

Comparisons of IR boundaries and sequence divergence analysis

Among the 42 *Glycine* plastomes, IR regions range from 25,405 to 25,581 bp. At the four IR-LSC/SSC junctions, the LSC-IR_B boundary (JLB) resides within *rps*19, which spans 211–220 bp in the LSC and 59–68 bp in IR_B. The IR_B-SSC boundary (JSB) lies within *ndh*F in the subgenus *Glycine*, where *ndh*F extends 1–42 bp into IR_B and 2,234–2,262 bp into SSC, whereas in the subgenus *Soja* it coincides with the 3'-end of *ndh*F. The SSC-IR_A boundary (JSA) occurs within *ycf*1. In the SSC region, *ycf*1 spans 40 bp in *G. tomentella* (KC893633 and NC021636), 2,763 bp in *G. cyrtoloba* (NC021645), and 4,878–4,920 bp in the remaining 39 genomes, while consistently spanning 440–589 bp in IR_A. The IR_A-LSC boundary (JLA) lies between *rpl*2 and *trn*H(GUG), located 112–181 bp from *rpl*2 and 2–42 bp from *trn*H(GUG). Partial IR expansion or contraction was observed. Subgenus *Soja* exhibits IR expansion into both LSC and SSC regions. Within subgenus *Glycine* Clade A, *G. dolichocarpa*, *G. remota*, and *G. clandestina* (C_AA120811) show expansion toward the SSC region, while *G. latrobeana* displays expansion toward the LSC region. In Clade C, *G.* sp. (C_AA120819) exhibits IR contraction at the SSC boundary (Fig. S4).

Comparative analysis of 42 *Glycine* plastomes revealed higher sequence variation and nucleotide diversity in single-copy regions than in IR regions, and in non-coding regions than in coding regions (Fig. 4, S5). Nucleotide diversity (Pi) was < 0.01 in all non-coding sequences > 200 bp, while < 0.005 in protein-coding genes. Among protein-coding genes, *mat*K, *acc*D, *ycf*4, *cem*A, *rpl*20, and *rps*3 (located in LSC), as well as *ndh*F, *ccs*A, and *rps*15 (in SSC), showed elevated diversity (Pi > 0.010). In non-coding regions, the LSC-located *psb*A-*trn*H(GUG), *trn*G(UCC)-*trn*R(UCU), *pet*A-*psb*J, *trn*P(UGG)-*trn*W(CCA), *rpl*20-*rps*18, and *rps*19-*rps*3 regions exhibited high diversity (Pi > 0.03).

Repeat sequences identification

Among the 42 *Glycine* plastomes, 3,637 SSRs were identified, with 69% located in intergenic regions (Fig. 5B). The number of SSRs per plastome ranged from 75 in *G. curvata* to 102 in *G. falcata* (Fig. 5A; Table S4). Six SSR types were detected: mononucleotide repeats predominated (53.7%–69.4%, primarily A/T types), followed by dinucleotide (16.7%–27.6%), tetranucleotide (7.0%–17.3%), and trinucleotide repeats (1.2%–7.0%), while penta-/hexanucleotide repeats occurred only in some genomes (Fig. 5C).

A total of 3,655 long repeats were identified, including 1,127 forward, 593 reverse, 1,593 palindromic, and

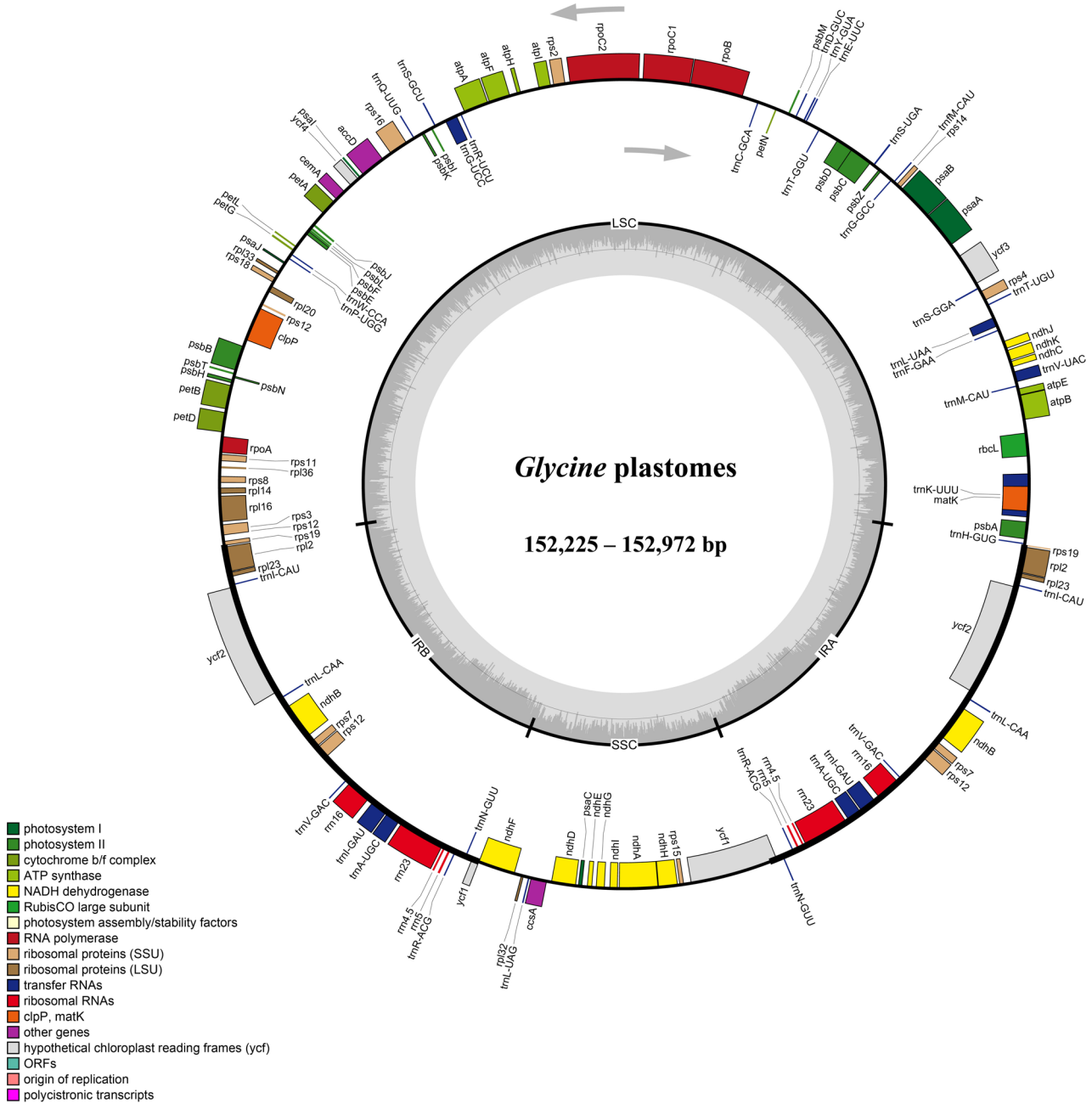

Fig. 2 Plastome map of *Glycine*. Genes inside the circle are transcribed clockwise while genes outside the circle are transcribed counterclockwise. Genes belonging to different functional groups are marked with different colors. The darker and lighter gray bands in the inner circle represent GC content and AT content, respectively

342 complementary repeats. All plastomes shared a conserved 287-bp forward repeat. The number of long repeats per plastome ranged from 60 in *G. clandestina* and *G. curvata* to 110 in *G. tomentella* (Fig. 5D; Table S5). Despite interspecific variation in number, distribution patterns remained consistent: palindromic repeats constituted the highest proportion (38.5%–53.4%), followed by forward (25.7%–36.5%), reverse (8.3%–21.2%), and complementary repeats (4.1%–15.3%). By length category, 30–45 bp repeats predominated (86.8%–94.8%), 45–60 bp repeats accounted for 3.1%–10.3%, >90 bp repeats comprised 1.0%–1.7%, while 60–75 bp and 75–90 bp repeats were scarce or absent (Fig. 5E).

Codon usage pattern

Across the 42 *Glycine* plastomes, protein-coding regions contained from 20,819 (*G. canescens*_C_AA120825) to 22,589 (*G. microphylla*_ C_AA120816) codons (Fig. S7).

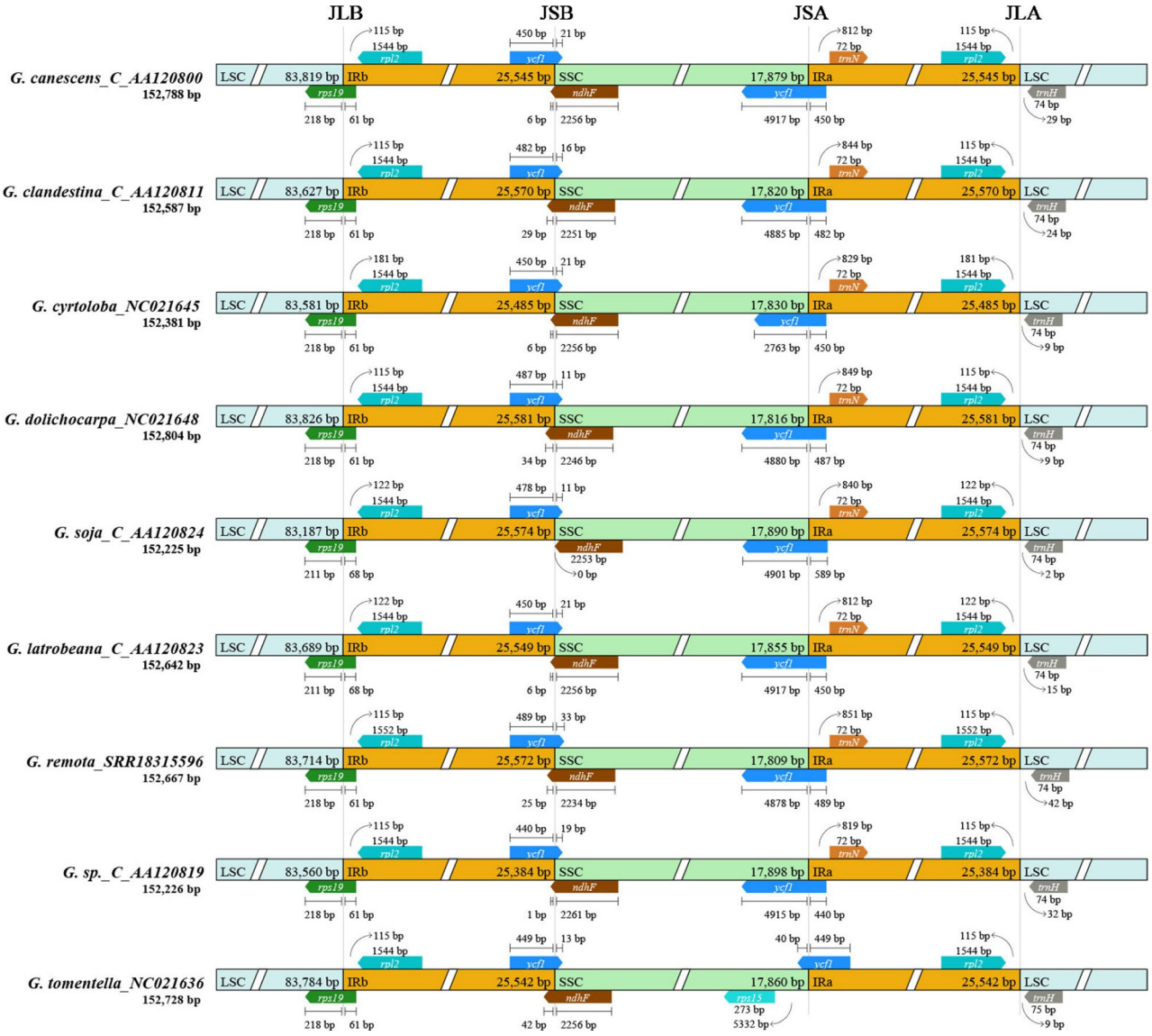

Fig. 3 Comparative analysis of IR boundary regions in nine representative samples selected from 42 *Glycine* plastomes revealed a conserved gene organization, with only minor expansions or contractions. These structural variations show no significant correlation with overall plastome size divergence

The most frequent codon was AUU (952–1,036 occurrences), encoding isoleucine. UUA (encoding leucine) exhibited the highest relative synonymous codon usage (RSCU) value (mean = 2.14). Thirty-one codons displayed RSCU > 1.0, with 30 terminating in A/U nucleotides. Both AUG (encoding methionine) and UGG (encoding tryptophan) consistently had RSCU = 1.0 across all genomes, indicating no codon usage bias (Fig. 6; Table S6).

Selective pressure analysis

Genes with invalid d_N/d_S values (ω = 0.0001, due to near 100% identity; ω = 999, due to d_S = 0) were considered statistical outliers and excluded from the calculation to prevent bias. After excluding 13 outliers, analysis of 64 plastid protein-coding genes across *Glycine* species revealed d_N/d_S ratios of 0.01718–0.66607, indicating pervasive purifying selection (Fig. 7). Branch-model analysis evaluated selection pressure on eight foreground branches (Fig. 1). Likelihood ratio tests showed the M0 model best fit most genes. After excluding aberrant d_N/d_S values (ω = 999 or ω = 0.0001), three genes exhibited lineage-specific selection shifts: *acc*D underwent relaxed purifying selection in subgenus *Soja* (Fig. 1, FG1); *rpo*B showed relaxed purifying selection in *G. falcata* (Fig. 1, FG5); and *rbc*L experienced significant positive selection within a monophyletic group comprising all species in clade A of subgenus *Glycine*, except *G. latrobeana* and *G. falcata* (Fig. 1, FG8).

Table 1 Protein-coding genes identified in the plastomes of 42 *Glycine* samples (15 species) categorized by function

Function of Genes	Group of Genes	Name of Genes
Self–replication	Ribosomal RNA genes	*rrn*4.5×2, *rrn*5×2, *rrn*16×2, *rrn*23×2
	Transfer RNA genes	*trn*A(UGC)* ×2, *trn*C(GCA), *trn*D(GUC), *trn*E(UUC), *trn*F(GAA), *trnf*M(CAU), *trn*G(GCC), *trn*G(UCC)*, *trn*H(GUG), *trn*I(CAU) ×2, *trn*I(GAU)* ×2, *trn*K(UUU)*, *trn*L(CAA) ×2, *trn*L(UAA)*, *trn*L(UAG), *trn*M(CAU), *trn*N(GUU) ×2, *trn*P(UGG), *trn*Q(UUG), *trn*R(ACG) ×2, *trn*R(UCU), *trn*S(GCU), *trn*S(GGA), *trn*S(UGA), *trn*T(GGU), *trn*T(UGU), *trn*V(GAC) ×2, *trn*V(UAC)*, *trn*W(CCA), *trn*Y(GUA)
	Ribosomal protein (small subunit)	*rps*2, *rps*3, *rps*4, *rps*7×2, *rps*8, *rps*11, *rps*12**×2, *rps*14, *rps*15, *rps*16*, *rps*18, *rps*19
	Ribosomal protein (large subunit)	*rpl*2*×2, *rpl*14, *rpl*16*, *rpl*20, *rpl*23×2, *rpl*32, *rpl*33, *rpl*36
	RNA polymerase	*rpo*A, *rpo*B, *rpo*C1*, *rpo*C2
Genes for photosynthesis	Subunits of photosystem I	*psa*A, *psa*B, *psa*C, *psa*I, *psa*J
	Subunits of photosystem II	*psb*A, *psb*B, *psb*C, *psb*D, *psb*E, *psb*F, *psb*H, *psb*I, *psb*J, *psb*K, *psb*L, *psb*M, *psb*N, *psb*T, *psb*Z
	Subunits of ATP synthase	*atp*A, *atp*B, *atp*E, *atp*F*, *atp*H, *atp*I
	Large subunit of Rubisco	*rbc*L
	Subunits of NADH Dehydrogenase	*ndh*A*, *ndh*B*×2, *ndh*C, *ndh*D, *ndh*E, *ndh*F, *ndh*G, *ndh*H, *ndh*I, *ndh*J, *ndh*K
	Cytochrome b6f complex	*pet*A, *pet*B*, *pet*D*, *pet*G, *pet*L, *pet*N
Other genes	Maturase	*mat*K
	Envelope membrane protein	*cem*A
	Subunit of acetyl–CoA	*acc*D
	Synthesis gene	*ccs*A
	ATP–dependent protease	*clp*P**
	Genes of unknown function	*ycf*1, *ycf*2×2, *ycf*3**, *ycf*4

Gene*:Gene with one introns; Gene**:Gene with two introns;Gene (×2):Number of copies of multi-copy genes

Discussion

Plastome characteristics and adaptive evolution of *Glycine*

The plastomes within the subfamily Papilionoideae generally exhibit pronounced structural variations. Although most lineages of this subfamily including *Glycine* share a common ~ 50 kb large inversion, plastomes of *Glycine* are highly conserved compared to other closely related genera. For instance, *Medicago* L. has undergone frequent inversions and rearrangements [67], while *Trifolium* L. is characterized by the loss of *acc*D and significant gene duplications [68].

A comparative analysis of 42 *Glycine* plastomes confirms this stability: their structure, size, gene content, and gene order are highly consistent, with minimal variation in total length (~ 0.8 kb). In angiosperms, plastome size fluctuation is often linked to the expansion or contraction of the Inverted Repeat (IR) regions. However, in *Glycine*, plastome size variation shows no apparent correlation with IR dynamics. The minor IR boundary shifts observed in *Glycine* are also documented in other genera such as *Indigofera* (Leguminosae) [69], *Salvia* (Lamiaceae) [70], and *Amygdalus* (Rosaceae) [71]; such subtle variations at the boundaries may provide valuable reference information for species delimitation within the genus. Given this high degree of macro-structural stability, genetic differentiation in *Glycine* is primarily reflected in micro-scale sequence variations. Consequently, an in-depth exploration of nucleotide diversity within these conserved genomes is significant for revealing mutational hotspots and evolutionary pressures across the genus.

Glycine plastomes were highly conserved in sequence, with IR regions showing lower nucleotide diversity than SC regions, and protein-coding regions lower than non-coding regions––patterns consistent with other angiosperms (e.g., *Cardiocrinum* (Endl.) Lindl [72]., *Quercus* L [73]., *Crataegus* L [74]., *Ficus* L [75]., and *Rubus* L [76]). Despite this, six divergent hotspot regions (Pi > 0.03) were identified in SC regions: *psb*A-*trn*H(GUG), *trn*G(UCC)-*trn*R(UCU), *pet*A-*psb*J, *trn*P(UGG)-*trn*W(CCA), *rpl*20-*rps*18, and *rps*19-*rps*3. Among these, *psb*A-*trn*H(GUG) and *pet*A-*psb*J have been widely used for phylogenetic reconstruction and species identification across many angiosperm groups [77, 78]. Research indicates that sequences that evolve at different rates vary in their utility for resolving phylogenetic relationships at different taxonomic levels. For instance, Wang et al. [79] demonstrated that intron sequences could resolve deep relationships within Fabaceae but were less effective at resolving intergeneric relationships within subfamilies. Conversely, intergenic spacers, characterized by faster evolutionary rates, proved more effective for resolving relationships among closely related taxa. Sakai et al. [29] utilized a concatenated matrix of four plastid intergenic spacers (*rps*11-*rpl*36, *atp*B-*rbc*L, *trn*S-*trn*G, *rps*3-*rpl*16) to construct a phylogeny for *Glycine* but failed to resolve some deep nodes. However, none of these four spacers were identified as hotspots in this study. Thus, the identified hotspots represent potential markers for phylogenetic reconstruction, population genetics, and DNA barcoding in *Glycine* and related taxa.

Glycine has undergone radiation and allopolyploidization [1], resulting in complex ploidy in species such as *G.*

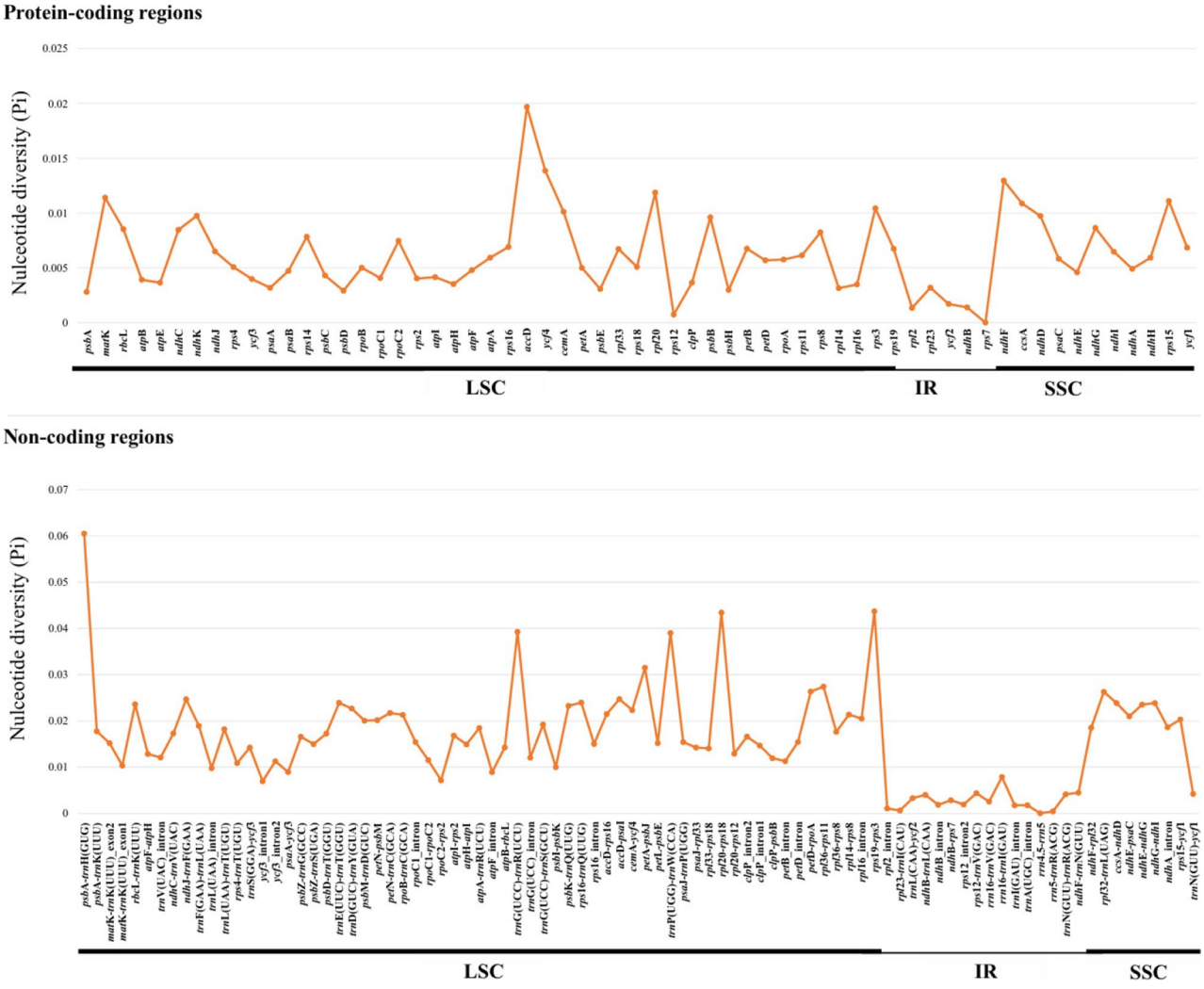

Fig. 4 Nucleotide diversity (Pi) values across 42 *Glycine* plastomes, showing genus-wide divergence patterns. The upper panel shows Pi values for protein-coding sequences, whereas the lower panel shows Pi values of non-coding regions

tomentella and *G. tabacina* [80]. The allopolyploidization hypothesis is supported by evidence from hybridization experiments [81], isozyme analysis [82], morphology [83], DNA data [84] and phylogenomic data [85]. These processes may have shaped significant selective imprints on the plastome. Selection pressure analysis revealed three genes under lineage-specific shifts: *rpo*B (relaxed selection in *G. falcata*), *acc*D (relaxed selection in Subgenus *Soja*), and *rbc*L (significant positive selection in foreground branch 8), suggesting that these genes may be associated with environmental adaptation processes within their respective lineages.

The *acc*D gene encodes a subunit of acetyl-CoA carboxylase (ACCase), a key enzyme in fatty acid synthesis that regulates unsaturated fatty acid production and influences biomembrane stability under high temperatures [86]. The editing efficiency of *acc*D correlates with thermotolerance. Wild soybean (*G. soja*) of the eastern Asian subgenus *Soja*, occurs widely across China, where northeastern China being a genetic diversity center. In contrast to species of the subgenus *Glycine*, subgenus *Soja* has undergone long-term adaptation to colder environments. Therefore, we hypothesize that the observed selection pressure on *acc*D in subgenus *Soja* may be linked to its low-temperature adaptation.

In *G. falcata*, *rpo*B (encoding the β subunit of the plastid-encoded RNA polymerase, PEP) is under relaxed selection [87]. PEP serves as the primary RNA polymerase in mature plastid, responsible for transcribing > 80% of plastid genes and directly regulates the expression of photosynthesis-related genes [88–90]; its activity is modulated by environmental signals like light, temperature, and circadian rhythms [87]. Furthermore, *rpo*B influences photosynthetic efficiency by coordinating nuclear-plastid gene interactions [91]. While *G. falcata* occupies a broad ecological range, our comparative

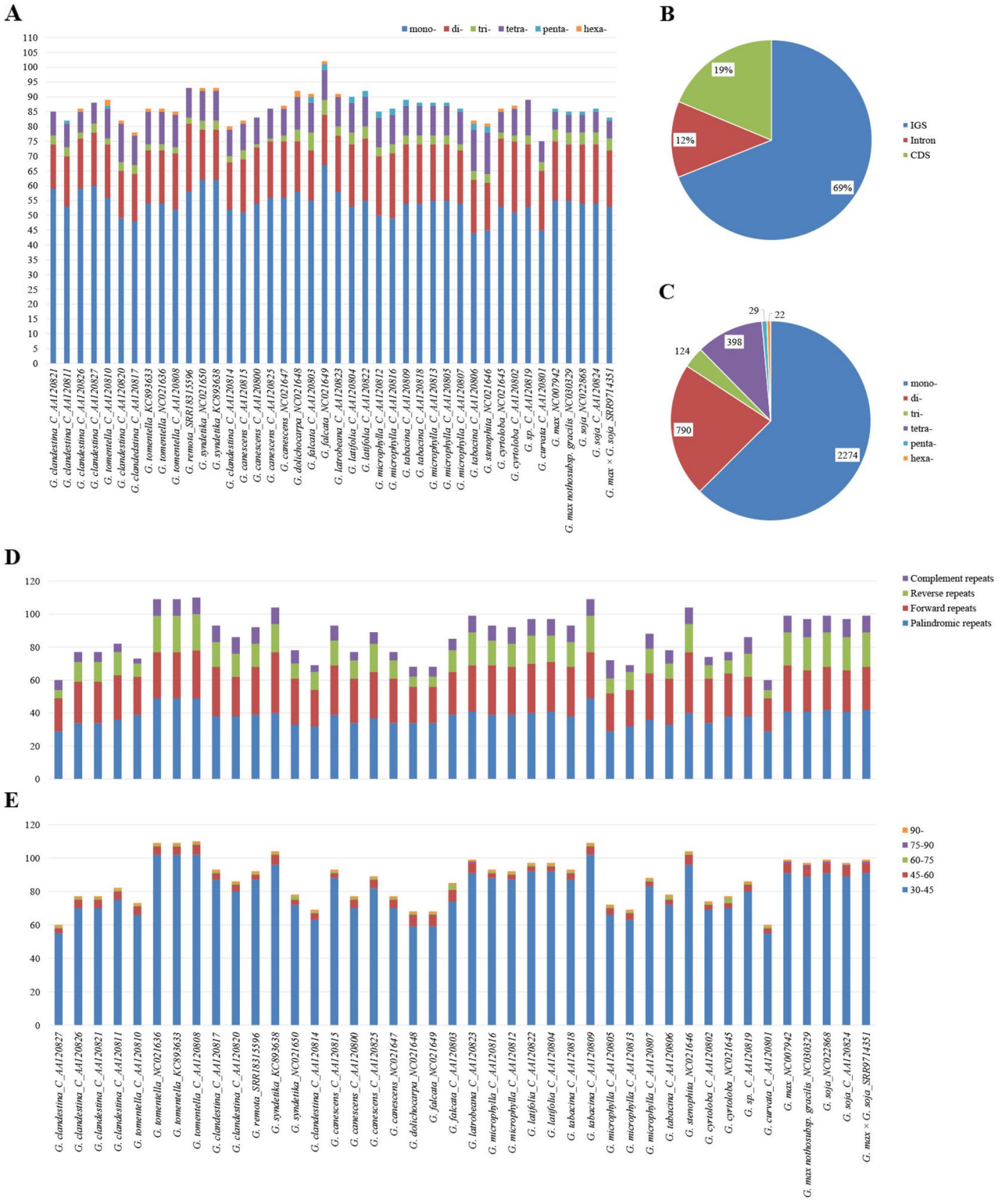

Fig. 5 Distribution of simple sequence repeats (SSRs) and long repeats sequences in 42 *Glycine* plastomes: (**A**) types and quantities of SSRs detected in each species; (**B**) distribution of SSRs across distinct regions of the plastomes; (**C**) abundance of different SSR types; (**D**) types and abundance of long repeat sequences identified in each species; (**E**) length intervals and corresponding counts of these sequences across studied species

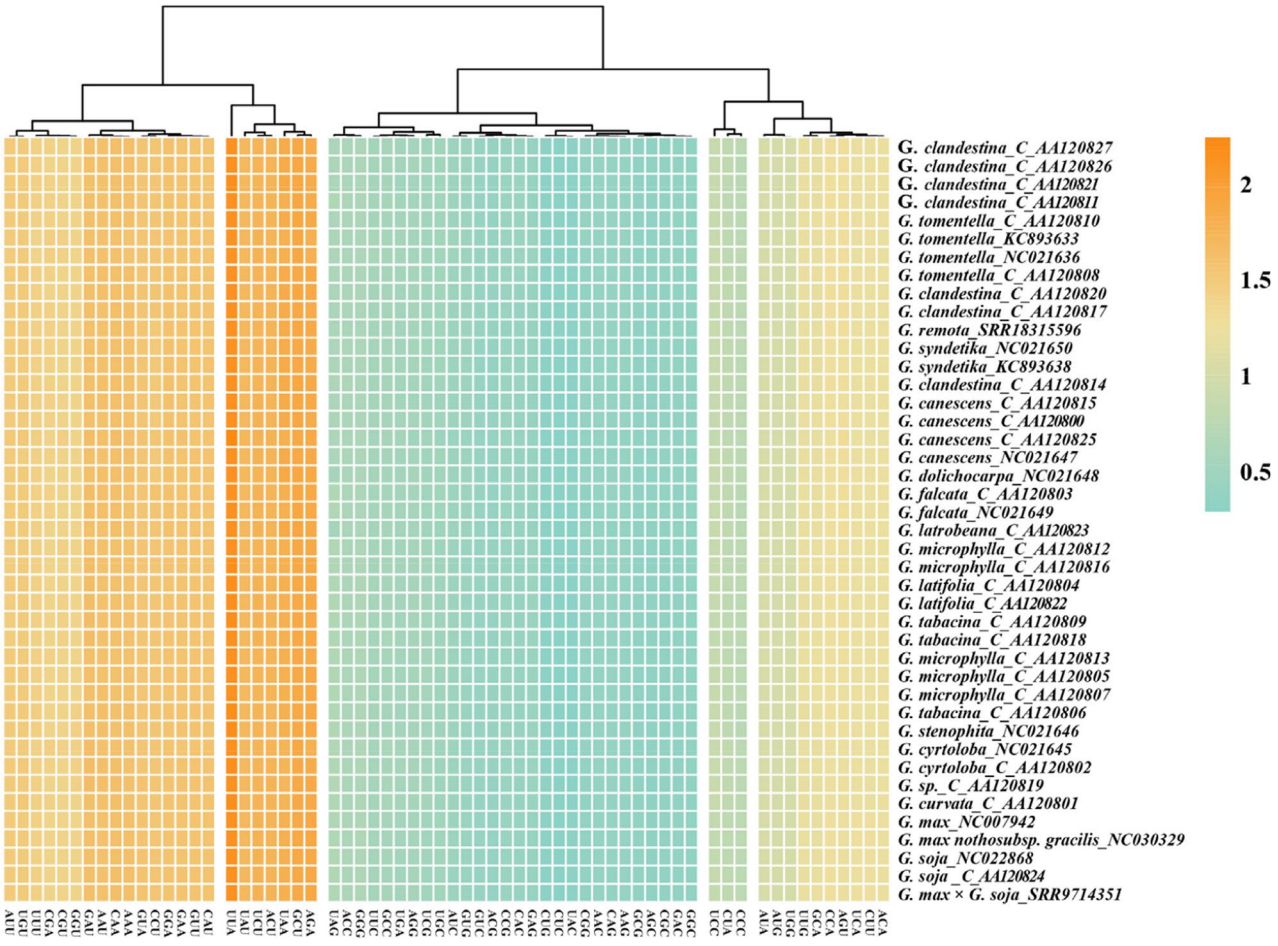

Fig. 6 Relative Synonymous Codon Usage (RSCU) patterns in plastid protein-coding sequences across 42 *Glycine* plastomes

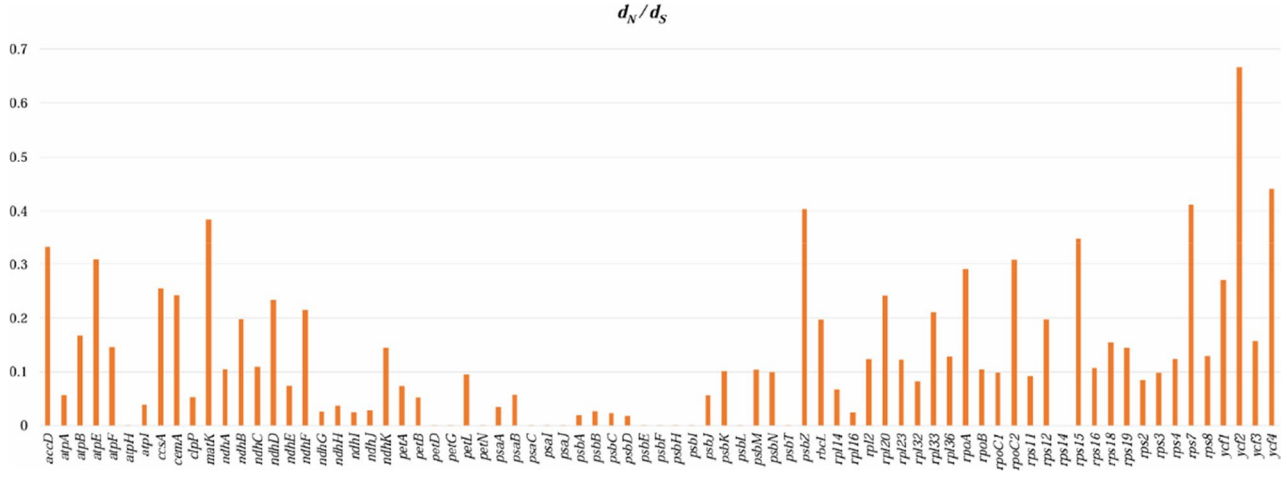

Fig. 7 Non-synonymous to synonymous substitution rate ratios (d_N/d_S) for all protein-coding genes in 42 *Glycine* plastomes

analysis with other widespread *Glycine* species (e.g., *G. tomentella* and *G. tabacina*) suggests that this relaxed selection is specific to *G. falcata* rather than a universal trait of widespread species. This specific evolutionary pattern may be associated with the unique ecological pressures or evolutionary history experienced by *G. falcata*.

Within foreground branch 8 (Fig. 1, FG8), the *rbc*L gene showed significant positive selection. The *rbc*L encodes the large subunit of Rubisco, which contains the

catalytic core and is responsible for the vital physiological process of CO_2 fixation in photosynthesis. Given that Rubisco's catalytic efficiency and CO_2/ O_2 specificity are highly sensitive to temperature and CO_2 availability, the evolution of its primary sequence is widely regarded as a critical mechanism for plant habitat adaptation. Similar patterns in *Ficus* [75] and *Ilex* [92] link positive selection on *rbc*L to ecological adaptability. The species in foreground branch 8 (*G. dolichocarpa*, *G. canescens*, *G. syndetika*, *G. remota*, *G. tomentella*, and *G. clandestina*) occupy diverse habitats from arid to humid regions suggesting that the positive selection at specific sites may enhance Rubisco's performance under the specific environmental stresses (e.g., high temperature or aridity) encountered during the radiation of this group, even though similar selection was not observed in all widespread *Glycine* species (e.g., *G. falcata* and *G. tabacina*).

Across the 42 *Glycine* plastomes, all 77 genes were under purifying selection ($d_N/d_S < 1$), consistent with patterns in other angiosperms [75, 76, 93–97]. Purifying selection maintains essential functional domains by purging deleterious mutations [98, 99], explaining the high interspecific conservation observed in *Glycine* plastomes.

Phylogenetic analysis

We present the most comprehensively sampled plastomic genome phylogeny of the genus *Glycine* to date, revealing a well-supported phylogeny across major clades. Consistent with previous studies [30, 31, 100], the monophyly of the genus *Glycine* and its two subgenera, *Soja* and *Glycine*, is strongly supported. Within the subgenus *Glycine*, three major clades (A, B, and C) are recovered. Clade B and Clade C form a clade that is sister to Clade A, consistent with plastome-based phylogenies [30, 31], but incongruent with nuclear genome-based phylogenies (e.g., histone H3-D in Sherman-Broyles et al. [30] and genome-wide SNP data in Landis and Doyle [31]), which place *G. falcata* as the earliest-diverging lineage. These clades, initially proposed by Doyle et al. [28], are here expanded. Clade A is expanded to include additional species such as *G. dolichocarpa*, *G. syndetika*, and *G. remota*, encompassing a substantial portion of the perennial *Glycine* diversity. It is widely distributed across diverse ecological zones in Australia, including deserts, dry shrublands, temperate grasslands, savannas, and temperate broadleaf and mixed forests. Clade B is extended to include *G. stenophita*. Species in Clade B are unified by morphological features such as stoloniferous growth, adventitious root buds (except *G. stenophita* and partial allotetraploid group of *G. tabacina* [81]) and specific secondary venation patterns of leaves [28], and some produce fertile hybrids in interspecific crosses [101–103]. Together, these morphological characteristics clarify the evolutionary coherence of Clade B. However, they remain insufficient to justify formal taxonomic revisions within this clade. Clade C contains two species characterized by distinctively curved pods. Due to differences in taxon sampling, interspecific relationships are difficult to compare directly across studies.

Our plastome phylogeny includes multiple accessions for 15 species, allowing finer-scale inference. While certain species remain unsampled due to their restricted distributions in remote geographic regions and the consequent challenge in obtaining high-quality molecular materials, our taxon sampling encompasses all recognized major clades within subgenus *Glycine* according to established phylogenetic frameworks. Therefore, the current sampling density is sufficient to represent the species-level evolutionary breadth of the genus. However, it is undeniable that insufficient species sampling within Clade A constrains the resolution of its internal phylogenetic relationships. Several species appear non-monophyletic, including *G. canescens*, *G. clandestina*, *G. microphylla*, *G. tomentella*, and *G. tabacina* (Fig. 1). Within Clade A, the plastid phylogeny shows *G. canescens* and *G. clandestina* to be non-monophyletic, contrasting with their monophyly in nuclear datasets [5, 31]. Such cytonuclear discordance is common in angiosperms (e.g., *Trema* [104], *Amorpha* [105]) and has been frequently documented in association with plastome capture events [106]. Within Clade B, *G. microphylla* is shown here, for the first time, to be non-monophyletic. Two accessions form a weakly supported clade with *G. latifolia*, with this clade being sister to *G. tabacina*; while three others cluster as a strongly supported sister group (BS = 100%; Fig. 1). This topology differs from previous studies based on ITS sequences and plastid non-coding fragments [27, 29] and suggests a possible complexity in its intraspecific relationships that warrants further investigation, especially considering that hybridization among *G. microphylla*, *G. latifolia*, and *G. tabacina* has been documented [77]. Non-monophyly in *G. tomentella* and *G. tabacina*, species representing classic polyploid complexes (2n = 38/40/78/80), has been consistently documented and reflects their allopolyploid origins [1]. Previous studies delineated *G. tomentella* into six diploid (D1–D6) and six tetraploid taxa (T1–T6) [82, 107], leading to partial taxonomic resolution (e.g., *G. arenaria* Tindale and *G. syndetika* split from *G. tomentella*, *G. stenophita* from *G. tabacina*) [108–110]. Nevertheless, while plastomic phylogeny offers crucial insights into the evolutionary history of *Glycine*, these data alone are insufficient to propose formal taxonomic revisions or species delimitations for these non-monophyletic taxa, and sampling limitations, given the extensive ranges of certain taxa, may introduce bias into phylogenetic inferences. Recent studies show the presence of multiple plastid lineages with *Macadamia* species due to

geographically based plastid capture events between species [21]. The divergence likely stems from differing plastome inheritance patterns across lineages, potentially driven by interspecific hybridization [81]. Resolving taxonomy in these cases requires population-level sampling and multilocus analyses. Future studies should prioritize broader intraspecific sampling and integrate nuclear phylogenomics to untangle the complex evolutionary past of this genus. Ultimately, such genomic data should be combined with morphological evidence to establish robust species boundaries within these intricate lineages.

The phylogenetic incongruence of *G. clandestina*, *G. latrobeana*, and *G. falcata* exemplifies the cytonuclear discordance within *Glycine*. This pattern is consistent with hybridization/introgression or incomplete lineage sorting (ILS) frequently observed in other Papilionoideae genera such as *Lotus* L [111]. and *Trifolium* [112], typically attributed to reticulate evolution. Overall, despite the highly conservation of the *Glycine* plastome structure, its evolutionary history is profoundly shaped by reticulate processes. Such discordance underscores that reticulate evolution is a pervasive driver of legume diversification. Consequently, single plastid datasets are insufficient to resolve the complex evolution of Fabaceae, necessitating integrative phylogenomic analyses with multi-omic data in future studies.

Glycine latrobeana is placed as sister to other Clade A species (excluding *G. falcata*), consistent with Doyle et al. [27] but conflicting with plastid fragment data [28], which grouped it with *G. clandestina* and *G. canescens*. This inconsistency may reflect both cytonuclear conflict and limited resolution from short plastid sequences. A clearer placement will require broader taxon sampling and multiple data types.

The placement of *G. falcata* also varies between nuclear and plastome datasets. Our plastome-based tree positions it within Clade A, divergent following *G. latrobeana* (BS = 82%), as sister to all other members, in agreement with recent plastome-based studies [29, 30] (these studies did not sample *G. latrobeana*). In contrast, nuclear phylogenies place *G. falcata* as sister to the entire subgenus *Glycine* (e.g. [5, 30]), . This cytonuclear discordance suggests a complex evolutionary history involving ancient hybridization or incomplete lineage sorting. Resolving *G. falcata*'s placement will require integration of nuclear genomes, transcriptomes, and multiple phylogenetic methods to disentangle the drivers of discordance and clarify the evolutionary relationships within the genus *Glycine*.

While this study achieved high overall phylogenetic resolution, support for certain nodes remains relatively low. This may reflect the rapid radiation experienced by *Glycine* during its early evolution, wherein limited phylogenetic signal (e.g., synapomorphies) accumulated over a short divergence time, approaching the intrinsic resolving limit of single plastid genomic datasets and making it difficult to reliably distinguish between incomplete lineage sorting (ILS) and hybridization events. Furthermore, due to insufficient geographic coverage in intraspecific sampling, the observed topological conflicts may not fully represent the genuine evolutionary history. Therefore, integrating nuclear genomic data with coalescent-based analytical approaches will be necessary to determine whether these patterns of non-monophyly and plastome capture reflect genuine evolutionary processes or stem from analytical bias due to insufficient informative sites.

Conclusion

In this study, we analyzed 42 plastomes (including 28 newly sequenced, 2 reassembled from public raw data and the remaining 12 from GenBank) representing 15 *Glycine* species. These plastomes exhibit the highly conserved quadripartite structure, gene content (111 genes: 77 PCGs, 30 tRNAs, 4 rRNAs) and strong collinearity across the genus. Six non-coding regions were found to be highly diverse, which can serve as candidate markers for future phylogenetics, population genomics, and species delimitation in the genus *Glycine*. Phylogenomic analyses provide robust support for the monophyly of the genus *Glycine* and its two subgenera (*Glycine* and *Soja*), and resolve three major clades (A, B, C) within subgenus *Glycine*, with Clades B and C forming a clade sister to Clade A. Notably, six species (*G. canescens*, *G. clandestina*, *G. microphylla*, *G. tabacina*, *G. tomentella*, and *G. soja*) exhibited non-monophyly across accessions, suggesting complex histories potentially involving hybridization. Our study further reveals cytonuclear discordance in the phylogeny of *Glycine*, indicating its complicated evolutionary history probably shaped by hybridization, plastid capture, and incomplete lineage sorting. Further studies should include population-level sampling of non-monophyletic species to delimit intraspecific lineages and clarify taxonomic boundaries by integrating molecular and morphological evidence, and implement comprehensive sampling across *Glycine* to reconstruct a robust phylogeny using phylogenomic datasets and elucidate the mechanisms underlying cytonuclear discordance.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12870-026-08422-z.

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements

We are grateful to Prof. Xiao-Hua Jin, the Chinese National Herbarium, Institute of Botany, Chinese Academy of Sciences (PE) and the Queensland

Herbarium (BRI) for their assistance in plant sampling. We also thank the Molecular Biology Experiment Center, The Germplasm Bank of Wild Species in Southwest China (Kunming Institute of Botany, CAS) for sequencing support. Furthermore, we extend our special thanks to Dr. Qin Tian for her insightful suggestions and constructive feedback.

Authors' contributions

K.-L.A.: Methodology, Formal analysis, Visualization, Writing. W.G.: Formal analysis, Visualization, Writing. Q.L.: Methodology, Visualization. D.-J.W.: Methodology, Formal analysis. Z.X.: Methodology, Formal analysis. X.-G.F.: Methodology, Formal analysis. L.Z.: Formal analysis. Z.-Y.Y.: DNA extraction and genome sequencing. R.H.: review & editing. J.W.: Sample collection, review & editing. R.Z.: Sample collection and genome sequencing, Funding acquisition, Writing. T.-S.Y.: Conceptualization, Methodology, Writing, Funding acquisition. All authors approved the manuscript.

Funding

This research was supported by the Basic Research Project of Yunnan Province (No. 202401BC070001), Yunnan Revitalization Talent Support Program: Yunling Scholar Project (No. XDYC-YLXZ-2024-0021) and Young Talent Project (XDYC-QNRC-2022-0257), National Natural Science Foundation of China (No. 32270247), Yunnan Applied & Basic Research Program (No. 202301AT070310), Science and Technology Basic Resources Investigation Program of China (No. 2019FY100900), National Key Research and Development Program of China (No. 2022YFC2601202), the National Natural Science Foundation of China, and Key International (regional) Cooperative Research Project (No. 31720103903).

Data availability

All newly generated plastome data in this study are available in the National Genomics Data Center (https://ngdc.cncb.ac.cn/genbase) under accession numbers provided in shown in Supplemental Table S1.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 20 September 2025 / Accepted: 17 February 2026
Published online: 09 March 2026

References

1. Sherman-Broyles S, Bombarely A, Powell AF, Doyle JL, Egan AN, Coate JE, et al. The wild side of a major crop: Soybean's perennial cousins from Down Under. Am J Bot. 2014;101:1651–65.
2. Ren S, Gilbert MG. *Glycine* Willd. In: Wu ZY, Raven PH, Hong DY, editors. Flora of China(volume 10). Beijing and St. Louis: Science Press and Missouri Botanical Garden; 2010. pp. 250–52.
3. Bauer S, Hymowitz T, Noel G. Soybean cyst nematode resistance derived from *Glycine tomentella* in amphiploid (*G. max* × *G. tomentella*) hybrid lines. Nematropica. 2007;37:277–86.
4. Riggs RD, Wang S, Singh RJ, Hymowitz T. Possible transfer of resistance to Heterodera glycines from *Glycine tomentella* to *Glycine max*. J Nematol. 1998;30:547.
5. Hwang EY, Wei H, Schroeder SG, Fickus EW, Quigley CV, Elia P, et al. Genetic diversity and phylogenetic relationships of annual and perennial *Glycine* species. (Bethesda). 2019;G3:9:2325–36.
6. Linnaeus C. Species Plantarum. Stockholm: Laurentii Salvii; 1753.
7. Willdenow KL. Species Plantarum(volume 3). Third ed. Berlin: G. C. Nauk; 1802.
8. de Candolle AP. Mémoires sur la famille des Légumineuses. Paris: A. Belin; 1825.
9. Hermann FJ. A revision of the genus *Glycine* and its immediate allies. Washington, DC, USA: USDA; 1962.
10. Verdcourt B. A proposal concerning *Glycine* L. Taxon. 1966;15:34–6.
11. Lackey J. *Neonotonia*, a new generic name to include *Glycine wightii* (Arnott) Verdcourt (Leguminosae, Papilionoideae). Phytologia. 1977;37:209–12.
12. Soltis ED, Soltis PS. Contributions of plant molecular systematics to studies of molecular evolution. Plant Mol Biol. 2000;42:45–75.
13. Gitzendanner MA, Soltis PS, Yi TS, Li DZ, Soltis DE. Plastome phylogenetics: 30 years of inferences into plant evolution. Adv Bot Res. 2018;85:293–313.
14. Smith DR. Mutation rates in plastid genomes: They are lower than you might think. Genome Biol Evol. 2015;7:1227–34.
15. Daniell H, Jin S, Zhu XG, Gitzendanner MA, Soltis DE, Soltis PS. Green giant—a tiny chloroplast genome with mighty power to produce high-value proteins: History and phylogeny. Plant Biotechnol J. 2021;19:430–47.
16. Birky CW Jr. Uniparental inheritance of mitochondrial and chloroplast genes: Mechanisms and evolution. Proc Natl Acad Sci. 1995;92:11331–38.
17. Parks M, Cronn R, Liston A. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biol. 2009;7:1–17.
18. Barrett CF, Freudenstein JV, Li J, Mayfield-Jones DR, Perez L, Pires JC, et al. Investigating the path of plastid genome degradation in an early-transitional clade of heterotrophic orchids, and implications for heterotrophic angiosperms. Mol Biol Evol. 2014;31:3095–112.
19. Barrett CF, Davis JI, Leebens-Mack J, Conran JG, Stevenson DW. Plastid genomes and deep relationships among the commelinid monocot angiosperms. Cladistics. 2013;29:65–87.
20. Wicke S, Müller KF, DePamphilis CW, Quandt D, Bellot S, Schneeweiss GM. Mechanistic model of evolutionary rate variation en route to a nonphotosynthetic lifestyle in plants. Proc Natl Acad Sci. 2016;113:9045–50.
21. Manatunga SL, Furtado A, Topp B, Alam M, Mason PJ, Kharabian-Masouleh A, et al. Analysis of phylogenetic relationships in *Macadamia* shows evidence of extensive reticulate evolution. Front Plant Sci. 2024;15:1394244.
22. Wijesundara UK, Furtado A, Dillon NL, Masouleh AK, Henry RJ. Phylogenetic relationships in the genus *Mangifera* based on whole chloroplast genome and nuclear genome sequences. Trop Plants. 2024;3:e034.
23. Nakandala U, Furtado A, Smith MW, Williams DC, Henry RJ. Phylogenetic relationships among Australian native citrus species based upon complete chloroplast genomes and single copy nuclear genes. Trop Plants. 2023;2:21.
24. Healey A, Lee DJ, Furtado A, Henry RJ. Evidence of inter-sectional chloroplast capture in Corymbia among sections Torellianae and Maculatae. Aust J Bot. 2018;66:369–78.
25. Wojciechowski MF, Sanderson MJ, Steele KP, Liston A. Molecular phylogeny of the temperate herbaceous tribes of papilionoid legumes: A supertree approach. In: Herendeen PS, Bruneau A, editors. Advances in Legume Systematics (part 9). UK: Royal Botanic Gardens, Kew; 2000. pp. 277–98.
26. Doyle JJ, Kanazin V, Shoemaker RC. Phylogenetic utility of histone H3 intron sequences in the perennial relatives of soybean (*Glycine*: Leguminosae). Mol Phylogenet Evol. 1996;6:438–47.
27. Kollipara KP, Singh RJ, Hymowitz T. Phylogenetic and genomic relationships in the genus *Glycine* Willd. based on sequences from the ITS region of nuclear rDNA. Genome. 1997;40:57–68.
28. Doyle JJ, Doyle JL, Brown AHD. A chloroplast-DNA phylogeny of the wild perennial relatives of soybean (*Glycine* subgenus *Glycine*): congruence with morphological and crossing groups. Evolution. 1990;44:371–89.
29. Sakai M, Kanazawa A, Fujii A, Thseng FS, Abe J, Shimamoto Y. Phylogenetic relationships of the chloroplast genomes in the genus *Glycine* inferred from four intergenic spacer sequences. Plant Syst Evol. 2003;239:29–54.
30. Sherman-Broyles S, Bombarely A, Grimwood J, Schmutz J, Doyle J. Complete plastome sequences from *Glycine syndetika* and six additional perennial wild relatives of soybean. G3. (Bethesda). 2014;4:2023–33.
31. Landis JB, Doyle JJ. Genomic relationships of *Glycine remota*, a recently discovered perennial relative of soybean, within *Glycine*. Syst Bot. 2023;48:78–87.
32. Yang XC, He JX, Zhou MH, Bi CW, Kong JL, Wang J, et al. Genomic variation and evolutionary patterns in organelle genomes between annual and perennial *Glycine* species. BMC Plant Biol. 2025;25:353.
33. Zhuang YB, Wang XT, Li XC, Hu JM, Fan LC, Landis JB, et al. Phylogenomics of the genus *Glycine* sheds light on polyploid evolution and life-strategy transition. Nat Plants. 2022;8:233–44.
34. Jansen RK, Ruhlman TA. Plastid genomes of seed plants. In: Bock R, Knoop V, editors. Genomics of Chloroplasts and Mitochondria (volume 35). Advances in Photosynthesis and Respiration. Dordrecht: Springer; 2012.
35. Daniell H, Lin CS, Yu M, Chang WJ. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biol. 2016;17:1–29.

36. Weng ML, Ruhlman TA, Jansen RK. Expansion of inverted repeat does not decrease substitution rates in *Pelargonium* plastid genomes. New Phytol. 2017;214:842–51.
37. Gao CW, Li T, Zhao X, Wu CH, Zhang Q, Zhao XZ, et al. Comparative analysis of the chloroplast genomes of *Rosa* species and RNA editing analysis. BMC Plant Biol. 2023;23:318.
38. Liu F, Chen NS, Wang HS, Li JM, Wang J, Qu F. Novel insights into chloroplast genome evolution in the green macroalgal genus *Ulva* (Ulvophyceae, Chlorophyta). Front Plant Sci. 2023;14:1126175.
39. Cauz-Santos LA, da Costa ZP, Callot C, Cauet S, Zucchi MI, Bergès H, et al. A repertory of rearrangements and the loss of an inverted repeat region in *Passiflora* chloroplast genomes. Genome Biol Evol. 2020;12:1841–57.
40. Palmer JD, Osorio B, Aldrich J, Thompson WF. Chloroplast DNA evolution among legumes: Loss of a large inverted repeat occurred prior to other sequence rearrangements. Curr Genet. 1987;11:275–86.
41. Palmer JD, Singh GP, Pillay DTN. Structure and sequence evolution of three legume chloroplast DNAs. Mol Gen Genet. 1983;190:13–9.
42. Palmer JD, Thompson WF. Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. Cell. 1982;29:537–50.
43. Doyle JJ, Doyle JL. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical Bull. 1987;19:11–5.
44. Jin JJ, Yu WB, Yang JB, Song Y, dePamphilis CW, Yi TS, et al. GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol. 2020;21:1–31.
45. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: Interactive visualization of de novo genome assemblies. Bioinformatics. 2015;31:3350–52.
46. Qu XJ, Moore MJ, Li DZ, Yi TS. PGA: A software package for rapid, accurate, and flexible batch annotation of plastomes. Plant Methods. 2019;15:1–12.
47. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28:1647–49.
48. Greiner S, Lehwark P, Bock R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: Expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Res. 2019;47:W59–64.
49. Li H, Guo QQ, Xu L, Gao HD, Liu L, Zhou XY. CPJSdraw: Analysis and visualization of junction sites of chloroplast genomes. PeerJ. 2023;11:e15326.
50. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: Computational tools for comparative genomics. Nucleic Acids Res. 2004;32:W273–79.
51. Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, et al. DnaSP 6: DNA sequence polymorphism analysis of large data sets. Mol Biol Evol. 2017;34:3299–302.
52. Thiel T, Michalek W, Varshney RK, Graner A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L). Theor Appl Genet. 2003;106:411–22.
53. Beier S, Thiel T, Münch T, Scholz U, Mascher M. MISA-web: A web server for microsatellite prediction. Bioinformatics. 2017;33:2583–85.
54. Kurtz S, Schleiermacher C, REPuter. Fast computation of maximal repeats in complete genomes. Bioinformatics. 1999;15:426–27.
55. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R, REPuter. The manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 2001;29:4633–42.
56. Hejazi FA, Mohammadi P, Soorni A. Comparative chloroplast genomics of *Teucrium* species reveals genome evolution, phylogenetic relationships, and candidate molecular markers. Sci Rep. 2025;15:44318.
57. Gohar SD, Soorni A. Plastome diversity and phylogenomic analysis of *Satureja* (Lamiaceae): Uncovering evolutionary patterns and diagnostic markers. Planta. 2025;262:149.
58. Akrami AM, Meratian Esfahani S, Soorni A. Decoding the chloroplast genomes of five Iranian *Salvia* species: Insights into genomic structure, phylogenetic relationships, and molecular marker developmen. BMC Genomics. 2025;26:545.
59. Soorni A, Golchini MM. Complete chloroplast genome of Mentha aquatica reveals hypervariable regions and resolves phylogenetic position within the genus *Mentha*. Mol Biol Rep. 2025;52:677.
60. Zhang D, Gao FL, Jakovlić I, Zou H, Zhang J, Li WX, et al. PhyloSuite: An integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. Mol Ecol Resour. 2020;20:348–55.
61. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–13.
62. Yang ZH. PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24:1586–91.
63. Gao FL, Chen CJ, Arab DA, Du ZG, He YH, Ho SY. EasyCodeML: A visual tool for analysis of selection using CodeML. Ecol Evol. 2019;9:3891–98.
64. Zhang R, Wang YH, Jin JJ, Stull GW, Bruneau A, Cardoso D, et al. Exploration of plastid phylogenomic conflict yields new insights into the deep relationships of Leguminosae. Syst Biol. 2020;69:613–22.
65. Zhao YY, Zhang R, Jiang KW, Qi J, Hu Y, Guo J, et al. Nuclear phylotranscriptomics and phylogenomics support numerous polyploidization events and hypotheses for the evolution of rhizobial nitrogen-fixing symbiosis in Fabaceae. Mol Plant. 2021;14:748–73.
66. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–73.
67. Wu S, Chen JY, Li Y, Liu A, Li A, Yin M, et al. Extensive genomic rearrangements mediated by repetitive sequences in plastomes of *Medicago* and its relatives. BMC Plant Biol. 2021;21:421.
68. Sveinsson S, Cronk Q. Evolutionary origin of highly repetitive plastid genomes within the clover genus (*Trifolium*). BMC Evol Biol. 2014;14:228.
69. Zhou SM, Wang F, Yan SY, Zhu ZM, Gao XF, Zhao XL. Phylogenomics and plastome evolution of *Indigofera* (Fabaceae). Front Plant Sci. 2023;14:1186598.
70. Du Q, Yang HY, Zeng J, Chen ZE, Zhou JC, Sun SH, et al. Comparative genomics and phylogenetic analysis of the chloroplast genomes in three medicinal *Salvia* species for bioexploration. Int J Mol Sci. 2022;23:12080.
71. Du ZY, Lu K, Zhang K, He YM, Wang HT, Chai GQ, et al. The chloroplast genome of *Amygdalus* L.(Rosaceae) reveals the phylogenetic relationship and divergence time. BMC Genomics. 2021;22:1–16.
72. Lu RS, Li P, Qiu YX. The complete chloroplast genomes of three *Cardiocrinum* (Liliaceae) species: Comparative genomic and phylogenetic analyses. Front Plant Sci. 2017;7:2054.
73. Yin KQ, Zhang Y, Li YJ, Du FK. Different natural selection pressures on the *atpF* gene in evergreen sclerophyllous and deciduous oak species: Evidence from comparative analysis of the complete chloroplast genome of *Quercus aquifolioides* with other oak species. Int J Mol Sci. 2018;19:1042.
74. Wu LW, Cui YX, Wang Q, Xu ZC, Wang Y, Lin YL, et al. Identification and phylogenetic analysis of five *Crataegus* species (Rosaceae) based on complete chloroplast genomes. Planta. 2021;254:1–12.
75. Zhang ZR, Yang X, Li WY, Peng YQ, Gao J. Comparative chloroplast genome analysis of *Ficus* (Moraceae): Insight into adaptive evolution and mutational hotspot regions. Front Plant Sci. 2022;13:965335.
76. Lu Q, Tian Q, Gu W, Yang CX, Wang DJ, Yi TS. Comparative genomics on chloroplasts of *Rubus* (Rosaceae). Genomics. 2024;116:110845.
77. Li XW, Yang Y, Henry RJ, Rossetto M, Wang YT, Chen SL. Plant DNA barcoding: from gene to genome. Biol Rev. 2015;90:157–66.
78. Shaw J, Shafer HL, Leonard OR, Kovach MJ, Schorr M, Morris AB. Chloroplast DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in angiosperms: the tortoise and the hare IV. Am J Bot. 2014;101:1987–2004.
79. Wang YH, Wicke S, Wang H, Jin JJ, Chen SY, Zhang SD, et al. Plastid genome evolution in the early-diverging legume subfamily Cercidoideae (Fabaceae). Front Plant Sci. 2018;9:138.
80. Kole C. Wild crop relatives: genomics and breeding resources. Heidelberg: Springer Berlin; 2011.
81. Hymowitz T, Singh RJ, Kollipara KP. The genomes of the *Glycine*. In: Jules Janick, editor. New York: John Wiley and Sons. Plant Breeding Reviews. 1998;(16):289–318.
82. Doyle MJ, Grant JE, Brown AHD. Reproductive isolation between isozyme groups of *Glycine tomentella* (Leguminosae), and spontaneous doubling in their hybrids. Aust J Bot. 1986;34:523–35.
83. Costanza SH, Hymowitz T. Adventitious roots in Glycine subg. Glycine (Leguminosae): Morphological and taxonomic indicators of the B genome. Plant Syst Evol. 1987;158:37–46.
84. Doyle JJ, Doyle JL, Rauscher JT, Brown AHD. Diploid and polyploid reticulate evolution throughout the history of the perennial soybeans (*Glycine* subgenus *Glycine*). New Phytol. 2004;161:121–32.
85. Bombarely A, Coate JE, Doyle JJ. Mining transcriptomic data to study the origins and evolution of a plant allopolyploid complex. PeerJ. 2014;2:e391.
86. Huang C, Liu D, Li ZA, Molloy DP, Luo ZF, Su Y, et al. The PPR protein RARE1-mediated editing of chloroplast *accD* transcripts is required for fatty acid biosynthesis and heat tolerance in *Arabidopsis*. Plant Commun. 2023;4:100461.
87. Wu XX, Mu WH, Li F, Sun SY, Cui CJ, Kim C, et al. Cryo-EM structures of the plant plastid-encoded RNA polymerase. Cell. 2024;187:1127–44. e21.

88. Zhelyazkova P, Sharma CM, Förstner KU, Liere K, Vogel J, Börner T. The primary transcriptome of barley chloroplasts: Numerous noncoding RNAs and the dominating role of the plastid-encoded RNA polymerase. Plant Cell. 2012;24:123–36.
89. Börner T, Aleynikova AY, Zubo YO, Kusnetsov VV. Chloroplast RNA polymerases: Role in chloroplast biogenesis. Biochim Biophys Acta. 2015;1847:761–69.
90. Hajdukiewicz PT, Allison LA, Maliga P. The two RNA polymerases encoded by the nuclear and the plastid compartments transcribe distinct groups of genes in tobacco plastids. EMBO J. 1997;16:4041–8.
91. Tadini L, Jeran N, Peracchio C, Masiero S, Colombo M, Pesaresi P. The plastid transcription machinery and its coordination with the expression of nuclear genome: Plastid-Encoded Polymerase, Nuclear-Encoded Polymerase and the Genomes Uncoupled 1-mediated retrograde communication. Philos Trans R Soc Lond B Biol Sci. 2020;375:20190399.
92. Yao X, Tan YH, Yang JB, Wang Y, Corlett RT, Manen JF. Exceptionally high rates of positive selection on the *rbcL* gene in the genus *Ilex* (Aquifoliaceae). BMC Evol Biol. 2019;19:1–13.
93. Souza UJBd, Vitorino LC, Bessa LA, Silva FG. The complete plastid genome of *Artocarpus camansi*: A high degree of conservation of the plastome structure in the family Moraceae. Forests. 2020;11:1179.
94. Wang LY, Wang J, He CY, Zhang JG, Zeng YF. Characterization and comparison of chloroplast genomes from two sympatric *Hippophae* species (Elaeagnaceae). J Res (Harbin). 2021;32:307–18.
95. Yang JY, Chiang YC, Hsu TW, Kim SH, Pak JH, Kim SC. Characterization and comparative analysis among plastome sequences of eight endemic *Rubus* (Rosaceae) species in Taiwan. Sci Rep. 2021;11:1152.
96. Henriquez CL, Abdullah, Ahmed I, Carlsen MM, Zuluaga A, Croat TB, et al. Molecular evolution of chloroplast genomes in Monsteroideae (Araceae). Planta. 2020;251:72.
97. Yang X, Xie DF, Chen JP, Zhou SD, Yu Y, He XJ. Comparative analysis of the complete chloroplast genomes in *Allium* Subgenus *Cyathophora* (Amaryllidaceae): Phylogenetic relationship and adaptive evolution. Biomed Res Int. 2020;2020:1732586.
98. Wu Z, Liao R, Yang T, Dong X, Lan D, Qin R, et al. Analysis of six chloroplast genomes provides insight into the evolution of *Chrysosplenium* (Saxifragaceae). BMC Genomics. 2020;21:1–14.
99. Huang R, Xie XN, Chen AM, Li F, Tian EW, Chao Z. The chloroplast genomes of four *Bupleurum* (Apiaceae) species endemic to Southwestern China, a diversity center of the genus, as well as their evolutionary implications and phylogenetic inferences. BMC Genomics. 2021;22:1–15.
100. Asaf S, Khan AL, Aaqil Khan M, Muhammad Imran Q, Kang SM, Al-Hosni K, et al. Comparative analysis of complete plastid genomes from wild soybean (*Glycine soja*) and nine other *Glycine* species. PLoS ONE. 2017;12:e0182281.
101. Putievsky E, Broué P. Cytogenetics of hybrids among perennial species of *Glycine* subgenus *Glycine*. Aust J Bot. 1979;27:713–23.
102. Newell CA, Hymowitz T. Hybridization in the genus *Glycine* subgenus *Glycine* willd.(Leguminosae, Papilionoideae). Am J Bot. 1983;70:334–48.
103. Grant JE, Grace JP, Brown AHD, Putievsky E. Interspecific hybridization in *Glycine* Willd. subgenus *Glycine* (Leguminosae). Aust J Bot. 1984;32:655–63.
104. Fu XG, Liu SY, van Velzen R, Stull GW, Tian Q, Li YX, et al. Phylogenomic analysis of the hemp family (Cannabaceae) reveals deep cyto-nuclear discordance and provides new insights into generic relationships. J Syst Evol. 2023;61:806–26.
105. Straub SC, Doyle JJ. Molecular phylogenetics of *Amorpha* (Fabaceae): An evaluation of monophyly, species relationships, and polyploid origins. Mol Phylogenet Evol. 2014;76:49–66.
106. Yi TS, Jin GH, Wen J. Chloroplast capture and intra-and inter-continental biogeographic diversification in the Asian–New World disjunct plant genus *Osmorhiza* (Apiaceae). Mol Phylogenet Evol. 2015;85:10–21.
107. Doyle MJ, Brown AHD. Numerical analysis of isozyme variation in *Glycine tomentella*. Biochem Syst Ecol. 1985;13:413–19.
108. Tindale MD. A new north Queensland species of *Glycine* Willd.(Fabaceae). Brunonia. 1986;9:99–103.
109. Pfeil BE, Craven LA, Brown AHD, Murray BG, Doyle JJ. Three new species of northern Australian *Glycine* (Fabaceae, Phaseolae), G. gracei, *G. montis-douglas* and *G. syndetika*. Aust Syst Bot. 2006;19: 245–58.
110. Doyle JJ, Doyle JL, Brown AHD, Pfeil BE. Confirmation of shared and divergent genomes in the *Glycine tabacina* polyploid complex (Leguminosae) using histone H3-D sequences. Syst Bot. 2000;25:437–48.
111. Kramina TE, Degtjareva GV, Samigullin TH, Valiejo-Romanet CM, Kirkbrideal JH, Volis S, et al. Phylogeny of Lotus (Leguminosae: Loteae): Partial incongruence between nrITS, nrETS and plastid markers and biogeographic implications. Taxon. 2016;65:997–1018.
112. Watson LE, Sayed-Ahmed H, Badr A. Molecular phylogeny of Old World Trifolium (Fabaceae), based on plastid and nuclear markers. Plant Syst Evol. 2000;224:153–71.

Publisher's note