Comparative analysis of complete plastid genomes from wild soybean (*Glycine soja*) and nine other *Glycine* species

**Sajjad Asaf¹, Abdul Latif Khan², Muhammad Aaqil Khan¹, Qari Muhammad Imran¹, Sang-Mo Kang¹, Khadija Al-Hosni¹, Eun Ju Jeong¹, Ko Eun Lee¹, In-Jung Lee¹**

¹ School of Applied Biosciences, Kyungpook National University, Daegu, Republic of Korea, ² Chair of Oman’s Medicinal Plants & Marine Natural Products, University of Nizwa, Nizwa, Oman

*ijlee@knu.ac.kr*

## Abstract

The plastid genomes of different plant species exhibit significant variation, thereby providing valuable markers for exploring evolutionary relationships and population genetics. *Glycine soja* (wild soybean) is recognized as the wild ancestor of cultivated soybean (*G. max*), representing a valuable genetic resource for soybean breeding programmes. In the present study, the complete plastid genome of *G. soja* was sequenced using Illumina paired-end sequencing and then compared it for the first time with previously reported plastid genome sequences from nine other *Glycine* species. The *G. soja* plastid genome was 152,224 bp in length and possessed a typical quadripartite structure, consisting of a pair of inverted repeats (IRa/IRb; 25,574 bp) separated by small (178,963 bp) and large (83,181 bp) single-copy regions, with a 51-kb inversion in the large single-copy region. The genome encoded 134 genes, including 87 protein-coding genes, eight ribosomal RNA genes, and 39 transfer RNA genes, and possessed 204 randomly distributed microsatellites, including 15 forward, 25 tandem, and 34 palindromic repeats. Whole-plastid genome comparisons revealed an overall high degree of sequence similarity between *G. max* and *G. gracilis* and some divergence in the intergenic spacers of other species. Greater numbers of indels and SNP substitutions were observed compared with *G. cyrtoloba*. The sequence of the *accD* gene from *G. soja* was highly divergent from those of the other species except for *G. max* and *G. gracilis*. Phylogenomic analyses of the complete plastid genomes and 76 shared genes yielded an identical topology and indicated that *G. soja* is closely related to *G. max* and *G. gracilis*. The complete *G. soja* genome sequenced in the present study is a valuable resource for investigating the population and evolutionary genetics of *Glycine* species and can be used to identify related species.

## Introduction

The chloroplast (cp) is a key organelle in photosynthesis and in the biosynthesis of fatty acids, starches, amino acids, and pigments [1, 2]. In angiosperms, plastomes are typically...
circular and highly conserved, ranging from 115 to 165 kb in length and comprising a small single-copy region (SSC; 16–27 kb) and a large single-copy region (LSC; 80–90 kb), separated by a pair of inverted repeats (IRs) [3, 4]. Most plastomes also contain 110–130 genes encoding up to 80 unique proteins and approximately 4 rRNAs and 30 tRNAs. Most of the protein-coding genes are associated with photosynthesis or other biochemical processes in plant cells, such as synthesis of amino acids, sugars, vitamins, lipids, pigments, and starches, storage, nitrogen metabolism, sulphate reduction, and immune responses [5, 6]. In contrast to mitochondrial and nuclear genomes, the plastomes of plants are highly conserved in regard to gene structure, organization, and content [4]. However, gene duplications, mutations, rearrangements, and losses have been observed in some angiosperm lineages [7]. Rearrangements of plastid gene order are generally observed in taxa with plastomes that exhibit at least one of the following qualities: variable IR region size, loss of one IR region, a high frequency of small dispersed repeats, complete or near-complete lack of photosynthesis, or biparental cp inheritance [8]. In addition, plastome inversions have been reported in a number of angiosperm families, including Asteraceae [9], Campanulaceae [10], Onagraceae [11], Leguminosae [12], and Geraniaceae [13, 14]. The plastomes of several members of the Papilionoideae also exhibit significant variation and rearrangement, including the loss of an IR region [15] and inversion of a 50-kb portion of the LSC [16, 17]. These features, as well as the loss of introns from the \textit{rps}12 and \textit{clpP} genes [18, 19] and transfer of \textit{rpl}22 to the nucleus [20, 21], have been well documented, and their occurrence has been mapped onto the phylogeny of Leguminosae [19].

The genus \textit{Glycine} comprises at least 28 species, which are separated into two subgenera, \textit{Glycine} and \textit{Soja}. The annuals include cultivated soybean, \textit{G. max}, and the wild soybean, \textit{G. soja}, that are native to eastern Asia, whereas most of the other species are perennials that are native to Australia. Researchers previously classified \textit{Glycine} species into various groups (A-I) on the basis of fertility of artificially produced hybrids and the degree to which meiotic chromosomes pair [22], and Ratnaparkhe et al. (2011) [23] further reviewed the nine genome groups using isozymes and sequences of two nuclear loci (H3D and nrDNA ITS).

Plastid data from various \textit{Glycine} species (annual and perennial) have been used in studies of phylogenetic and genetic diversity [24–28], including the investigation of neopolyploidy [29, 30]. For example, Doyle et al. (1990b) [24] identified three major clades within the perennial subgenus, showing varying degrees of agreement with nuclear phylogenies. However, additional research revealed incongruence between the plastid and nuclear phylogenies of the various genome groups [31]. The most noticeable incongruity was the placement of \textit{G. falcata}, which is the sole species in the F-genome group. According to the H3D gene-based phylogeny, \textit{G. falcata} is sister to all other perennial species, whereas chloroplast DNA fragment-based phylogenies strongly supported the placement of \textit{G. falcata} in the A-genome clade [16, 30, 32].

The advent of high-throughput sequencing technology has facilitated rapid progress in the field of genomics, especially in cp genetics. Since the first plastome was sequenced in 1986 [33], over 800 complete plastid genome sequences have been made available through the National Center for Biotechnology Information (NCBI) organelle genome database, including 300 from crop and tree genomes [34]. To date, complete plastomes have been reported for nine \textit{Glycine} species [35–37]. In the present study, the complete plastome of \textit{G. soja} was sequenced (GenBank accession number: KY241814) with the aim of elucidating global patterns of structural variation in the \textit{G. soja} plastome and comparing it for the first time with the available plastomes of nine other \textit{Glycine} species (\textit{G. max}, \textit{G. gracilis}, \textit{G. canescens}, \textit{G. cyrtoloba}, \textit{G. dolichocarpa}, \textit{G. falcata}, \textit{G. stenophita}, \textit{G. syndetika}, and \textit{G. tomentella}).

\textbf{Competing interests:} The authors have no competing interest.
Materials and methods
Chloroplast genome sequencing and assembly
The *G. soja* (accession KLG90379), seeds were received from the National Gene Bank of the Rural Development Administration of the Republic of Korea. Plants were cultivated in greenhouse at the Kyungpook National University, Republic of Korea. Plastid DNA was extracted from young leaves using the protocol described by Hu et al. [38], and the resulting DNA was sequenced using the Illumina HiSeq-2000 platform (San Diego, CA, USA) at Macrogen (Seoul, Korea). The *G. soja* plastome was then assembled *de novo* using a bioinformatics pipeline (http://phyzen.com). More specifically, a 400-bp paired-end library was produced according to the Illumina PE standard protocol, which resulted in 28,110,596 bp of sequence data, with a 101-bp average read length. Raw reads with Phred scores of 20 or lower were removed from the total PE reads using the CLC-quality trim tool, and *de novo* assembly of the trimmed reads was accomplished using CLC Genomics Workbench v7.0 (CLC Bio, Aarhus, Denmark) with a minimum overlap of 200 to 600 bp. The resulting contigs were compared against the *G. max* plastome using BLASTN with an E-value cutoff of 1e-5, and five contigs were identified and temporarily arranged based on their mapping position in the reference genome. After initial assembly, primers were designed (S1 Table) based on the terminal sequences of adjacent contigs, and PCR amplification and subsequent DNA sequencing were employed to fill in the gaps. PCR amplification was performed in 20-μl reactions that contained 1× reaction buffer, 0.4 μl dNTPs (10 mM), 0.1 μl Taq (Solg h-Taq DNA Polymerase), 1 μl (10 pm/μl) primers, and 1 μl (10 ng/μl) DNA, under the following conditions: initial denaturation at 95˚C for 5 min; 35 cycles of 95˚C for 30 s, 60˚C for 20 s, and 72˚C for 30 s; and a final extension step of 72˚C for 5 min. After incorporating the additional sequencing results, the complete plastome was used as a reference to map the remaining unmapped short reads to improve the sequence coverage of the assembled genome.

Analysis of gene content and sequence architecture
The *G. soja* plastome was annotated using DOGMA [39] and checked manually, and codon positions were adjusted based on comparison with homologs in the plastome of *G. max*. The transfer RNA sequences of the *G. soja* plastome were verified using tRNAscan-SE version 1.21 [40], with the default settings, and structural features were illustrated using OGDRAW [41]. To examine deviations in synonymous codon usage by avoiding the influence of the amino acid composition, the relative synonymous codon usage (RSCU) was determined using MEGA 6 [42]. Finally, the divergence of the new *G. soja* plastome from both perennial and annual *Glycine* species was assessed with mVISTA [43] in Shuffle-LAGAN mode, employing the new *G. soja* genome as a reference.

Characterization of repeat sequences and simple sequence repeats (SSRs)
Repeat sequences, including direct, reverse, and palindromic repeats, were identified within the plastome using REPuter [44], with the following settings: Hamming distance of 3, ≥90% sequence identity, and minimum repeat size of 30 bp. Additionally, SSRs were detected using Phobos version 3.3.12 [45], with the search parameters set to ≥10 repeat units for mononucleotide repeats, ≥8 repeat units for dinucleotide repeats, ≥4 repeat units for trinucleotide and tetranucleotide repeats, and ≥3 repeat units for pentanucleotide and hexanucleotide repeats. Tandem repeats were identified using Tandem Repeats Finder version 4.07 b [46], with default settings.
Sequence divergence and phylogenetic analyses

The average pairwise sequence divergence of 76 shared genes and the complete plastomes of 11 Glycine species were analysed using data from G. soja new (KY241814), G. soja old (NC022868), G. max, G. gracilis, G. canescens, G. cyrtoloba, G. dolichocarpa, G. falcata, G. stenophita, G. syndetika, and G. tomentella. Missing and ambiguous gene annotations were confirmed through comparative sequence analysis, after assembling a multiple sequence alignment and comparing gene order. The complete genome dataset was aligned using MAFFT version 7.222 [47], with default parameters, and Kimura’s two-parameter (K2P) model was selected to calculate pairwise sequence divergence [48]. A sliding window analysis was conducted to determine the nucleotide diversity (Pi) of the cp genome using DnaSP (DNA Sequences Polymorphism version 5.10.01) software [49]. The step size was set to 200 bp, with a window length of 800 bp. Similarly, Indel polymorphisms among the complete genomes were identified using DnaSP 5.10.01 [49], and a custom Python script (https://www.biostars.org/p/119214/) was employed to identify single-nucleotide polymorphisms. To resolve the phylogenetic position of G. soja within the genus Glycine, ten published Glycine species plastomes were downloaded from the NCBI database for phylogenetic analysis. Multiple alignment of the complete plastomes were constructed based on the conserved structure and gene order of the plastid genomes [8], and four methods were employed to construct phylogenetic trees: Bayesian inference (BI), implemented using MrBayes 3.1.2 [50]; maximum parsimony (MP), implemented using PAUP 4.0 [51]; and both maximum likelihood (ML) and joining-joining (NJ), implemented using MEGA 6 [42], employing previously described settings [52, 53]. In a second phylogenetic analysis, 76 shared cp genes from eleven Glycine species and two outgroup species (Phaseolus vulgaris and Vigna radiata) were aligned using ClustalX with default settings, followed by manual adjustment to preserve reading frames. Finally, the same four phylogenetic inference methods were employed to infer trees from the 76 concatenated genes, using the same settings [52, 53].

Results and discussion

Plastid genome organization

A total of 2,611,513 reads with an average read length of 101 bp were obtained, and these reads provided 1514.9× coverage of the plastome. The consensus sequence for a specific position was generated by assembling reads that were mapped with at least 934 reads per position and was used to construct the complete sequence of the G. soja plastome. The assembled G. soja plastome of was typical of angiosperms, with a pair of IR regions (25,574 bp), an LSC of 83,181 bp, and an SSC of 178,963 bp (Fig 1); a total size of 152,224 bp; and a GC content of 35.4% (Table 1). In addition, approximately 33.23% of the genome was non-coding, whereas protein-coding, rRNA, and tRNA genes constituted 52.06, 5.94, and 1.92% of the plastome, respectively (Table 2), similar to the values observed in other legume genomes. As observed in other angiosperm plastomes, the GC content was unequally distributed in the G. soja plastome; it was high in the IR regions (41.8%), moderate in the LSC region (32.8%), and low in the SSC region (28.73%; Table 1). The high GC content of the IR regions is due to the presence of eight ribosomal RNA (rRNA) sequences in these regions, as reported previously [54, 55].

The total coding DNA sequences (CDSs) were 79,250 bp in length and encoded 87 genes, including 26,416 codons (Table 3). The codon-usage frequency of the G. soja plastome was determined based on tRNA and protein-coding gene sequences (Table 4). Leucine (10.6%) and cysteine (1.2%) were the most and least frequently encoded amino acids, respectively, and
isoleucine, serine, glycine, arginine, and alanine constituted 9.0%, 7.7%, 6.5%, 5.8%, and 5.0% of the CDSs, respectively, as reported previously [54, 56].

Among these codons, the most and least frequently used were AAA (n = 1,181), which encodes lysine, and ATC and ATT (n = 1, n = 1), which both encode methionine. The AT contents of the 1st, 2nd, and 3rd codon positions of CDSs were 55.7%, 62.9%, and 72.4%, respectively (Table 3). The high AT content observed at the 3rd codon position is similar to that
reported for the plastomes of other terrestrial plants [54, 57, 58]. In addition, 46.36% and 57.65% of the preferred synonymous codons (RSCU > 1) ended with A or U and C or G, respectively, and 44.30% and 55.20% of the non-preferred synonymous codons (RSCU < 1) ended with C or G and A or U, respectively. However, there was no bias in start codon usage (AUG or UGG; RSCU = 1; Table 4).

Table 1. Summary of complete chloroplas t genomes for ten Glycine species.

| Region | G. soja a | G. soja | G. max | G. gracil | G. canes | G. cyrtol | G. doli | G. falca | G. stenop | G. syndet | G. tome |
|--------|-----------|---------|--------|-----------|----------|-----------|--------|---------|-----------|-----------|--------|
| LSC    | 83,181    | 83,174  | 83,174 | 83,175    | 83,579   | 83,174    | 83,815 | 84,027  | 83,937    | 83,839    | 83,773 |
| GC(%)  | 32.8      | 32.8    | 32.8   | 32.8      | 32.7     | 32.7      | 32.7   | 32.7    | 32.8      | 32.7      | 32.7   |
| Length | 54.64     | 54.64   | 54.64  | 54.64      | 54.58    | 54.8      | 54.9   | 54.99   | 54.87     | 54.85     | 54.85  |
| SSC    | 17,896    | 17,895  | 17,896 | 17,895     | 17,880   | 17,838    | 17,807 | 17,817  | 17,817    | 17,817    | 17,829 |
| GC(%)  | 28.7      | 28.8    | 28.8   | 28.8      | 28.6     | 28.6      | 28.7   | 28.7    | 28.8      | 28.7      | 28.7   |
| Length | 11.75     | 11.75   | 11.75  | 11.75      | 11.70    | 11.65     | 11.66  | 11.67   | 11.68     | 16.67     | 16.67  |
| IR     | 25,574    | 25,574  | 25,574 | 25,574     | 25,530   | 25,485    | 25,465 | 25,432  | 25,542     | 25,563     |       |
| GC(%)  | 41.8      | 41.9    | 41.9   | 41.9      | 41.9     | 41.9      | 41.9   | 41.8    | 41.9      | 41.9      | 41.9   |
| Length | 16.80     | 16.8    | 16.8   | 16.8      | 16.77    | 16.72     | 16.74  | 16.66   | 16.71     | 16.71     | 16.73  |
| Total  | 35.4      | 35.4    | 35.4   | 35.4      | 35.3     | 35.3      | 35.3   | 35.3    | 35.3      | 35.3      | 35.3   |
| Length | 152,224   | 152,217 | 152,218| 152,218    | 152,218  | 152,218   | 152,381| 152,804 | 153,023   | 152,618   | 152,783|

G. soja a = G. soja new (in this study),
G. soja = G. soja (old), G. gracil = G. gracilis, G. canes = G. canescens, G. cyrtol = G. cyrtoloba, G. doli = G. dolichocarpa, G. falca = G. falcata, G. stenop = G. stenophila, G. syndet = G. syndetika, G. tome = G. tomentella

https://doi.org/10.1371/journal.pone.0182281.t001

Table 2. Comparison of coding and non-coding region size among ten Glycine species.

| Region | G. soja a | G. soja | G. max | G. gracil | G. canes | G. cyrtol | G. doli | G. falca | G. stenop | G. syndet | G. tome |
|--------|-----------|---------|--------|-----------|----------|-----------|--------|---------|-----------|-----------|--------|
| Protein Coding | | | | | | | | | | |
| Length (bp) | 79,250 | 77,835  | 77,769 | 77,811    | 77,607   | 72,294    | 77,649 | 77,598  | 77,646    | 77,604    | 77,601 |
| GC(%)  | 36.2      | 36.1    | 36.1   | 36.1      | 36.1     | 36.1      | 36.1   | 36.1    | 36.1      | 36.1      | 36.1   |
| Length (%) | 52.06     | 51.13   | 51.12  | 51.11     | 50.98    | 47.44     | 50.91  | 50.71   | 50.8      | 50.7      | 50.8   |
| rRNA    | 2,925     | 2,817   | 2,799  | 2,799     | 2,792    | 2,792     | 2,792  | 2,792   | 2,792     | 2,792     | 2,792 |
| GC(%)  | 52.4      | 52.9    | 52.9   | 53.0      | 52.8     | 52.8      | 52.8   | 52.8    | 52.8      | 52.8      | 52.8   |
| Length (%) | 1.92     | 1.85    | 1.83   | 1.83      | 1.83     | 1.82      | 1.82   | 1.82    | 1.82      | 1.82      | 1.82 |
| Intergenic | | | | | | | | | | |
| Length (bp) | 60,995   | 62,511  | 62,603 | 62,554    | 62,765   | 68,241    | 63,309 | 63,579  | 63,123    | 63,333    | 63,281 |

G. soja a = G. soja (in this study),
G. soja = G. soja (old), G. max = G. max, G. gracil = G. gracilis, G. canes = G. canescens, G. cyrtol = G. cyrtoloba, G. doli = G. dolichocarpa, G. falca = G. falcata, G. stenop = G. stenophila, G. syndet = G. syndetika, G. tome = G. tomentella

https://doi.org/10.1371/journal.pone.0182281.t002
### Table 3. Base composition of the *G. soja* plastid genome.

|                   | T/U(%) | C (%) | A (%) | G(%) | Length (bp) |
|-------------------|--------|-------|-------|------|-------------|
| Genome            | 32.3   | 17.4  | 32.4  | 18.0 | 152,224     |
| LSC               | 33.6   | 16.0  | 33.6  | 16.8 | 83,181      |
| SSC               | 35.3   | 13.6  | 36.0  | 15.1 | 17,896      |
| IR                | 29.0   | 21.7  | 29.2  | 20.1 | 25,574      |
| rRNA              | 25.2   | 23.1  | 22.4  | 29.3 | 2,925       |
| Protein-coding genes | 32.2 | 17.0  | 31.5  | 19.2 | 79,250      |
| 1st position      | 24.1   | 18.3  | 31.6  | 25.8 | 26,416      |
| 2nd position      | 33.2   | 19.8  | 29.7  | 17.1 | 26,416      |
| 3rd position      | 39.2   | 12.7  | 33.2  | 14.7 | 26,416      |

[https://doi.org/10.1371/journal.pone.0182281.t003](https://doi.org/10.1371/journal.pone.0182281.t003)

### Table 4. The codon-anticodon recognition pattern and codon usage for the *G. soja* plastid genome.

| Amino acid | Codon No RSCU | tRNA Anticodon | Codon No RSCU | tRNA Anticodon |
|------------|---------------|----------------|---------------|----------------|
| Phe        | UUU 1099 1.28 | Ala GCA        | 395 1.18 trnA-UGC |
| Phe        | UUC 503 0.7   | tmF-GAA        | 122 0.5        |
| Leu        | UUA 932 1.9   | tmL-UAA tRNA   | 846 1.5        |
| Leu        | UUG 557 1.1   | tmL-CAA tRNA   | 165 0.47 trnY-GUA tRNA |
| Leu        | CUU 589 1.29  | Stop UAG       | 1 0.74         |
| Leu        | CUC 172 0.4   | Stop UGA       | 0 0.80         |
| Leu        | CUA 381 0.87  | tmL-UAG tRNA   | 5 1.44         |
| Leu        | CUG 164 0.32  | His CAU        | 503 1.49       |
| Ile        | AUU 1170 1.51 | His CAC        | 134 0.50 trnH-GUG tRNA |
| Ile        | AUC 392 0.5   | tmL-GAU tRNA   | 764 1.53 trnQ-UUG tRNA |
| Ile        | AUU 827 0.89  | Gln CAG        | 200 0.49       |
| Met        | AUG 499 1     | tmM-CAU tRNA   | 1045 1.44      |
| Val        | GUU 533 1.50  | Asn AAC        | 286 0.55 trnQ-UUG tRNA |
| Val        | GUC 158 0.46  | tmV-GAC tRNA   | 1181 1.44 trnK-UUU tRNA |
| Val        | GUA 534 1.47  | tmV-UAC tRNA   | 331 0.55       |
| Val        | GUG 173 0.54  | Glu GAG        | 827 1.55       |
| Ser        | UGU 1150 1.56 | Asp GAC        | 204 0.44 trmD-GUC tRNA |
| Ser        | UCC 298 1.23  | tmS-GGA tRNA   | 1042 1.48 trnE-UUC tRNA |
| Ser        | UCA 442 1.03  | tmS-UGA tRNA   | 313 0.51       |
| Ser        | UCG 181 0.48  | Cys UGU        | 231 1.50       |
| Ser        | AGU 405 1.24  | Cys UGC        | 85 0.49        |
| Ser        | AGC 120 0.42  | tmS-GCU tRNA   | 442 1 trmW-CCA tRNA |
| Pro        | CCU 403 1.59  | Arg CGU        | 339 1.36 trmR-ACG tRNA |
| Pro        | CCC 202 0.86  | Arg CGC        | 91 0.51        |
| Pro        | CCA 334 1.07  | tmP-UGG tRNA   | 361 1.24       |
| Pro        | CCG 122 0.47  | Arg CGG        | 100 0.48       |
| Thr        | ACU 571 1.68  | Arg AGA        | 485 1.77 trmR-UCU tRNA |
| Thr        | ACC 210 0.76  | tmT-GGU tRNA   | 156 0.61       |
| Thr        | ACA 421 1.08  | tmT-UGU tRNA   | 585 1.28       |
| Thr        | ACG 139 0.45  | Gly GGC        | 157 0.42       |
| Ala        | GCU 623 1.72  | Gly GGA        | 691 1.52 trnG-UCC tRNA |
| Ala        | GCC 189 0.59  | Gly GGG        | 282 0.77       |

[https://doi.org/10.1371/journal.pone.0182281.t004](https://doi.org/10.1371/journal.pone.0182281.t004)
Table 5. Genes in the sequenced G. soja chloroplast genome.

| Category            | Group of genes                  | Name of genes |
|---------------------|---------------------------------|----------------|
| **Self-replication**| Large subunit of ribosomal proteins | rpl2, 14, 16, 20, 22, 23, 32, 33, 36 |
|                     | Small subunit of ribosomal proteins | rps2, 3, 4, 7, 8, 11, 12, 14, 15, 16, 18, 19 |
|                     | DNA dependent RNA polymerase     | rpoA, B, C1, C2 |
| **tRNA genes**      |                                 | trnA-UGC, trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnM-CAU, trnG-UCC, trnH-GUG, trnL-CAU, trnL-GAU, trnK-UUU, trnL-CAG, trnL-UAG, trnM-CAU, trnN-GUU, trnP-GGG, trnP-UGG, trnQ-UUG, trnR-ACG, trnR-UCU, trnS-GCU, trnS-GGA, trnS-UGA, trnT-GGU, trnT-UGU, trnV-GAC, trnV-UAC, trnW-CCA, trnY-GUA |
| **Photosynthesis**   | Photosystem I                    | psaA, B, C, I, J |
|                     | Photosystem II                   | psbA, B, C, D, E, F, G, H, I, J, K, L, M, N, T, Z |
|                     | NADH oxidoreductase              | ndhA, B, C, D, E, F, G, H, I, J, K |
|                     | Cytochrome b6/f complex          | petA, B, D, G, L, N |
|                     | ATP synthase                     | atpA, B, E, F, H, I |
|                     | Rubisco                          | rbcL |
| **Other genes**     | Maturase                         | matK |
|                     | Protease                         | clpP |
|                     | Envelop membrane protein         | cemA |
|                     | Subunit Acetyl-CoA-Carboxylate    | accD |
|                     | c-type cytochrome synthesis gene | ccsA |
| **Unknown**         | Conserved Open reading frames    | ycf1, 2, 3, 15 |

https://doi.org/10.1371/journal.pone.0182281.t005

The G. soja genome map (Fig 1) was representative of known Glycine plastomes in general, and no structural rearrangement was detected among these plastomes. The length of the G. soja plastome was 152,224 bp, which is similar to that of G. max (152,217 bp) [35], but smaller than those of G. dolichocarpa, G. falkata, G. sydetika, and G. tomentella (Table 1). Among the sequenced Glycine plastomes, that of G. max is smallest, and that of G. dolichocarpa is largest (Table 1). Furthermore, a total of 134 genes were identified in the G. soja plastome, of which 110 were unique, including 87 protein-coding genes, 39 tRNA genes, and eight rRNA genes (Fig 1, Table 5). Similar to other legumes, the plastome of G. soja lacked the rpl22 gene, probably due to an ancient transfer to the nuclear genome [39]. The duplicated IR regions of the G. soja plastome resulted in complete duplication of the rpl2, rpl23, ycf2, ycf15, ndhB, and rps7 genes as well as duplication of exons 1 and 2 of rps12, all four rRNA genes, and seven tRNA genes. The LSC region included 61 protein-coding and 24 tRNA genes, whereas the SSC region included only 12 protein-coding genes and one tRNA gene. The protein-coding genes included nine genes encoding large ribosomal proteins (rpl2, 14, 16, 20, 22, 23, 32, 33, and 36), 12 genes encoding small ribosomal proteins (rps2, 3, 4, 7, 8, 11, 12, 14, 15, 16, 18, and 19), five genes encoding photosystem I components (psaA, B, C, I, and f), 16 genes related to photosystem II (Table 5), and six genes encoding ATP synthase and electron transport chain components (atpA, B, E, F, H, and I; Table 5).

Among the coding genes, rps12 was unequally divided, with its 5’ exon being located in the LSC region and one copy of the 3’ exon and intron being located in each of the IR regions, as in other angiosperms. The ycf1 gene was located at the IRa/SSC boundary, leading to incomplete duplication of the gene within the IR regions. We also identified 12 intron-containing
genes, including nine that contained a single intron and three (ycf3, clpP, and rps12) that contained two introns (Table 6). This is in contrast to the situation in Cicer arietinum, Medicago truncatula, Trifolium subterraneum, Pisum sativum, and Lathyrus sativus [19]. The largest intron was found in trnK-UUU (2583 bp) and included the entire matK gene, whereas trnL-UAA contained the smallest intron (508 bp).

Introns play an important role in the regulation of gene expression, and recent research has shown that introns can improve exogenous gene expression when located at specific positions. Therefore, introns can be a valuable tool for improving transformational efficiency [60]. Furthermore, intron sequences in legume chloroplast DNA have become important tools in phylogenetic analyses [61]. In addition, even though ycf1 and ycf2 [62, 63], rpl23 [64], and accD [65, 66] are often absent in plants [64], they have been reported to occur the plastomes of various Glycine species [67]. atpB-atpE pairs were observed to overlap with each other by ~1 bp. However, psbC-psbD exhibited a 53-bp overlap in G. soja plastomes, similar to what is observed in G. max [35] and G. falcata [67], Arabidopsis arenosa (17-bp overlap) [68], Gossypium (53-bp overlap) [69], and Camellia (52-bp overlap) [70]. Previously, Addachi et al. (2012) [71] reported the importance of the partial overlap of psbC and psbD cistrons. They demonstrated that the translation of the psbC cistron largely depends on the translation of the preceding psbD cistron, indicating a contribution form independent psbC translation. Similar results were reported in tobacco, where ndhC and ndhK cistrons overlap, and ndhK translation is strictly dependent on the upstream termination codon [72].

Repeat sequence content

Repeat analysis of the G. soja plastome identified 34 palindromic repeats, 15 forward repeats, and 25 tandem repeats (Fig 2A). Among these repeats, 12 of the forward repeats were 30–44

---

Table 6. Length of exons and introns in intron-containing genes from the *Glycine soja* plastid genome.

| Gene    | Location | Exon I (bp) | Intron I (bp) | Exon II (bp) | Intron II (bp) | Exon III (bp) |
|---------|----------|-------------|---------------|--------------|----------------|---------------|
| atpF    | LSC      | 144         | 736           | 414          |                 |               |
| clpP    | LSC      | 69          | 710           | 297          | 775            | 225           |
| ndhA    | SSC      | 552         | 1269          | 756          |                 |               |
| ndhB*   | IR       | 777         | 692           | 756          |                 |               |
| petB    | LSC      | 6           | 808           | 642          |                 |               |
| petD    | LSC      | 8           | 728           | 476          |                 |               |
| rpl2a   | IR       | 393         | 681           | 468          |                 |               |
| rpl16   | LSC      | 9           | 1165          | 402          |                 |               |
| rpoC1   | LSC      | 441         | 785           | 1638         | 719            | 159           |
| rps12*  | LSC      | 114         | -             | 26           | 531            | 232           |
| rps16   | LSC      | 39          | 887           | 228          |                 |               |
| ycf3    | LSC      | 126         | 697           | 228          | 745            | 150           |
| trnA-UGC| IR       | 38          | 810           | 35           |                 |               |
| trnL-GAU| IR       | 42          | 948           | 35           |                 |               |
| trnL-UAA| LSC      | 37          | 508           | 50           |                 |               |
| trnK-UUU| LSC      | 37          | 2583          | 29           |                 |               |
| trnV-UAC| LSC      | 39          | 586           | 37           |                 |               |

* replicated genes
*The rps12 coding sequence is split between 5′-rps12 and 3′-rps12, which are located in the large single-copy region and inverted repeat region, respectively.

https://doi.org/10.1371/journal.pone.0182281.t006
bp in length, while all 25 tandem repeats were 15–29 bp in length (Fig 2A–2D). Similarly, 27 of the palindromic repeats were 30–44 bp in length, and three repeats were 45–59 bp in length (Fig 2D). Overall, 74 repeats were identified in the *G. soja* plastome, which is a similar number to the 75, 75, 76, 83, 81, 83, 88, 80, and 80 repeat sequences found in the plastomes of *G. max*, *G. gracilis*, *G. canescens*, *G. cyrtoloba*, *G. dolichocarpa*, *G. falcata*, *G. stenophita*, *G. syndetika*, and *G. tomentella*, respectively (Fig 2A). Therefore, *G. soja* is more similar to *G. max* and *G. gracilis* in terms of repeats. Approximately 29.4% of these repeats were distributed in protein-coding regions. Previous reports suggest that repeat sequences, which contribute to genome rearrangements, can be very helpful in phylogenetic studies [58, 73]. In addition, analyses of various plastomes have shown that repeat sequences induce indels and substitutions [74], and both sequence variation and genome rearrangement occur as a result of slipped-strand mispairing and improper recombination of such repeat sequences [73, 75, 76]. Furthermore, the presence of repeat sequences indicates that loci are hotspots for genome reconfiguration.

Fig 2. Analysis of repeated sequences in 10 *Glycine* plastid genomes. A, Total of three repeat types; B, Length distribution of forward repeat sequences; C, Length distribution of tandem repeat sequences; D, Length distribution of palindromic repeat sequences.

https://doi.org/10.1371/journal.pone.0182281.g002
and repeats can be used to develop genetic markers for phylogenetic and population studies [58].

**SSR content**

Simple sequence repeats (SSRs), or microsatellites, are repeating sequences, typically of 1–6 bp in length, that are distributed throughout the genome. In the present study, we identified perfect SSRs in the plastome of *G. soja* and in those of nine other *Glycine* species (Fig 3A). Certain parameters were set because SSRs of 10 bp or longer are prone to slipped-strand mispairing, which is believed to be the main mechanism of the formation of SSR polymorphisms [78–80].

A total of 204 perfect microsatellites were identified in the *G. soja* plastome (Fig 3A), which is a similar number to the 206, 206, 210, 204, 216, 223, 213, 214, and 211 perfect microsatellites identified in the plastomes of *G. max*, *G. gracilis*, *G. canescens*, *G. cyrtoloba*, *G. dolichocarpa*, *G. falcata*, *G. stenophita*, *G. syndetika*, and *G. tomentella*, respectively (Fig 3A). The majority of the SSRs possessed dinucleotide repeat motifs, varying in number from 66 in *G. soja* to 76 in *G. falcata* and *G. dolichocarpa*, while trinucleotide SSRs were the second most common, ranging in number from 69 in *G. syndetika* to 74 in *G. stenophita*. Using our search criterion, two pentanucleotide SSRs were identified in *G. soja*, *G. max*, and *G. stenophita*,...
and two hexanucleotide SSRs were identified in G. gracilis and G. dolichocarpa (Fig 3A). In G. soja, the majority of the mononucleotide SSRs were A (98.1%) and C (1.81%) motifs, and the majority dinucleotide SSRs were A/T (71.64%) and A/G (23.94%) motifs (Fig 3B, Table 7). In addition, 61.7% of the SSRs were located in non-coding regions, whereas 2.9% and 0.49% were located in rRNA and tRNA genes, respectively (Fig 3C). Further analysis indicated that 64.7% of the SSRs were located in the LSC region, whereas 20.58% and 14.7% were located in the IR and SSC regions, respectively (Fig 3D). These results are similar to previous reports that SSRs are unevenly distributed in plastomes, and the findings might provide more information for selecting effective molecular markers for detecting intra- and interspecific polymorphisms [81–84]. Furthermore, most of the mono- and dinucleotide repeats consisted of A and T, which may have contributed to the bias in base composition, as in the plastomes of other species [85]. Our findings are comparable to previous reports that SSRs in plastomes are generally composed of polythymine (polyT) or polyadenine (polyA) repeats and infrequently contain tandem cytosine (C) or guanine (G) repeats [86], thereby contributing to AT richness [55, 56, 86].

Sequence and structural divergence of Glycine plastid genomes

Ten complete Glycine plastomes were compared with the G. soja plastome. Analysis of genes with known functions indicated that G. soja shared 76 protein-coding genes with nine Glycine species. In addition, the gene content and organization of the G. soja plastome were similar to those of other Glycine species plastomes [67], but different from the usual gene order of angiosperm plastomes, due to a large inversion (~51 kb) that reversed the order of the genes between trnK and accD (Fig 1). This 51-kb inversion was previously reported in other members of the legume family, especially members of subfamily Papilionoideae [16, 24, 87], and other inversions have been reported in the plastomes of other species, including a 5.6-kb inversion in Milletia [88], a 78-kb inversion in various closely related legumes, including Phaseolus and Vigna [17, 89], and a 36-kb inversion within the 51-kb inversion found in Lupinus and other genistoides [90]. This change in gene order has been ascribed to the contraction and expansion of IR regions, leaving the gene order as described in papilionoids, retaining the 51-kb inversion, but alerting the genes bordering the IR region [89, 91].

Furthermore, the IR region overlaps the ycf1 gene by 478 bp, as observed in legumes exhibiting the same inverted repeat as G. soja. This feature has been shown to distinguish the plastomes of legumes from those of other angiosperms, in which the IR region and ycf1 typically overlap by 1,000 bp [35]. Moreover, as found in the plastomes of other legumes, the plastome of G. soja possessed variation and was missing two cp genes, rpl22 and infA, [18], both of which have been replaced by cp-targeting nuclear copies [59, 92]. Absence of the rps16 gene from the plastome has also been reported in other legume lineages, excluding Glycine, and the mitochondrial copy is dually targeted to both the cp and mitochondria [19, 93]. Furthermore, loss of the introns in rps12 and clpP has been detected in the plastomes of various species [19], including those of Glycine species [35, 67].

Pairwise alignment of the new G. soja plastome with the old G. soja plastome and those of nine other genomes showed a high degree of synteny. The annotation of the new G. soja plastome was used as a reference for plotting the overall sequence identity of the plastomes of the other ten Glycine species in mVISTA (Fig 4). In the results, relatively lower sequence identity was observed between the plastomes of the seven other perennial species, especially in the rpoC1, atpF, accD, clpP, rpl2, ndhA, ndhF, rps8, rps19, and ycf1 genes (Fig 4). In addition, the LSC and SSC regions were less similar than the two IR regions in all Glycine species, and the non-coding regions were more divergent than the coding regions. Highly divergent regions
Table 7. Simple sequence repeats (SSRs) in the *Glycine soja* plastid genome.

| Unit | Length | No. | SSR start                   |
|------|--------|-----|------------------------------|
| A    | 18     | 1   | 51,531                      |
|      | 16     | 2   | 92,627, 142,764             |
|      | 15     | 2   | 76,538, 119,451             |
|      | 14     | 2   | 33,433, 82,862              |
|      | 13     | 4   | 24,610, 51,701, 110,244, 111,377 |
|      | 12     | 7   | 6,968, 9,644, 9,656, 58,365, 62,260, 75,661, 82,660 |
|      | 11     | 15  | 14,313, 42,712, 54,965, 59,329, 70,698, 78,955, 79,488, 81,034, 81,302, 10,9835, 111,046, 111,519, 111,927, 112,225, 122,146 |
|      | 10     | 22  | 2,391, 4,452, 7,568, 25,542, 31,495, 34,893, 38,160, 38,510, 45,234, 46,902, 54,259, 56,682, 62,419, 66,716, 67,450, 69,278, 93,297, 109,698, 110,547, 114,419, 124,220, 142,100 |
| C    | 12     | 1   | 9644                         |
| AT   | 19     | 1   | 5,177                        |
|      | 17     | 1   | 5,159                        |
|      | 16     | 1   | 24,676                       |
|      | 14     | 1   | 32,841                       |
|      | 13     | 1   | 48,415                       |
|      | 12     | 2   | 54,297, 118,666              |
|      | 11     | 8   | 33,695, 48,440, 65,081, 67,502, 68,320, 78,342, 79,508, 122,331 |
|      | 10     | 5   | 31,746, 32,806, 68,072, 80,714, 116,632 |
|      | 9      | 9   | 13,837, 35,671, 54,930, 58,400, 60,678, 64,792, 69,490, 82,699, 120,175 |
|      | 8      | 24  | 100, 1,607, 2,068, 3,635, 4,513, 4,526, 13,370, 16,835, 28,206, 47,399, 51,596, 51,773, 51,795, 58,249, 60,155, 65,092, 69,374, 76,625, 79,531, 82,378, 92,346, 116,291, 123,690, 143,053 |
| AG   | 9      | 2   | 25,492, 28,221               |
|      | 8      | 15  | 3,673, 6,261, 85,791, 86,793, 94,040, 105,226, 105,546, 107,047, 120,875, 128,352, 129,853, 130,173, 141,359, 148,606, 149,608 |
| AC   | 9      | 1   | 120,511                      |
| AAT  | 15     | 1   | 28,637                       |
|      | 13     | 1   | 14,614                       |
|      | 12     | 1   | 29,635                       |
|      | 11     | 1   | 73,972                       |
|      | 10     | 6   | 2,980, 14,647, 23,469, 47,482, 61,211, 83,153 |
|      | 9      | 15  | 4,840, 6,885, 18,582, 24,528, 28,614, 32,259, 32,318, 45,719, 47,151, 58,337, 80,973, 99,425, 115,619, 120,102, 135,973 |
| AAG  | 12     | 1   | 2,123                        |
|      | 11     | 1   | 111,544                      |
|      | 10     | 4   | 83,359, 95,785, 139,612, 152,038 |
|      | 9      | 15  | 23,601, 39,016, 61,479, 69,713, 76,888, 89,691, 91,515, 94,335, 102,444, 109,943, 117,624, 133,154, 141,063, 143,883, 145,707 |
| ATC  | 11     | 1   | 57,126                        |
|      | 9      | 6   | 22,369, 40,828, 45,626, 83,824, 116,434, 151,574 |
| ACG  | 10     | 2   | 83,313, 152,084              |
| AGC  | 9      | 5   | 5,366, 20,175, 68,568, 103,665, 131,733 |
| ACC  | 9      | 2   | 58,920, 90,061               |
| ACT  | 9      | 1   | 66,702                        |
| AGAT | 15     | 2   | 18,423, 18,450               |
| AATC | 13     | 1   | 119,923                      |
|      | 12     | 1   | 78,291                       |
| AAAG | 12     | 1   | 67,682                        |
| AAAT | 12     | 1   | 117,190                      |
| AACAG| 15     | 2   | 107,707, 127,685             |

https://doi.org/10.1371/journal.pone.0182281.t007
Fig 4. Visual alignment of plastid genomes from *Glycine soja* (new and old) and nine other *Glycine* species. VISTA-based identity plot showing the sequence identity among the ten *Glycine* species, using *G. soja* (new) as a reference. Vertical scale indicates the percentage of identity, ranging from 50% to 100%. Horizontal axis indicates the coordinates within the chloroplast genome. Arrows indicate the annotated genes and their transcriptional direction. A thick black line indicates the inverted repeat (IR) regions.

https://doi.org/10.1371/journal.pone.0182281.g004
The plastid genome comparison of *Glycine soja** included the *matK-rbcL*, *ycf3-psaA*, *trnC-rpoB*, *rpl20-clpP*, *rps16-trnQ*, *trnS-M-trnM*, *psbM-petN*, *atp1-atpH*, *petA-psbA*, and *ycf1-rps15* spacers, as reported previously [54, 55]. Our results also confirmed similar differences among various coding regions in the analysed species, as suggested by Kumar et al. [94]. On the other hand, *G. soja* exhibited high sequence identity with annual *Glycine* species (S1 Fig), which suggest that they are highly conserved. However, the variations in similarity levels revealed various coding and non-coding regions where the *G. soja* exhibits divergence from these annual *Glycine* species (S1 Fig). Similarly, we detected 10 relatively highly variable regions, including 4 gene regions and 6 intergenic regions of the cp genomes, that might be undergoing more rapid nucleotide substitution at species and cultivar levels (S2 Fig) (*atpB-rbcL*, *trnT-trnL*, *trnS(GGA)-trnG(UCC)*, *psbD-trnT*, *rps16*, *rpl13-rpl18*, *rpl16-rps3*, *ndhB*, *ycf1* and *ycf15*). These regions can be used as potential molecular markers for application in phylogenetic analyses of *Glycine*. Furthermore, various researchers have determined coding and non-coding regions of particularly high variability as potential molecular markers for *Glycine* species, such as *trnS(GGA)-trnG(UCC)*, *rpl16-rps3*, *trnT-trnL* and *atpB-rbcL* [95–97]. Similarly, it has been reported that non-coding regions in cp DNA show greater variability in nucleotide regions than coding regions, and these regions have become a major source of variability for phylogenetic studies in various species, including studies within *Glycine* species [98–100]. Furthermore, comparison of the plastomes of *G. soja* and related species revealed 72 SNPs and 26 indels in relation to *G. max* and *G. gracilis*, respectively (S2 Table). These results confirmed that the highly conserved plastome can include interspecific mutations that may be useful for analysing both genetic diversity and phylogenetic relationships.

Similarly, we calculated the average pairwise sequence divergence among the plastomes of the ten *Glycine* species (S3 Table). The plastome of *G. soja* exhibited an average sequence divergence of 0.0096, whereas that of *G. cyrtoloba* possessed the highest average sequence divergence (0.00567), and those of *G. soja* and *G. max* displayed the lowest average sequence divergence (0.00010 and 0.00020, respectively). Furthermore, the nine most divergent genes among these genomes were *accD*, *matK*, *ycf1*, *rps16*, *rpl20*, *psbM*, *psbN*, *petL*, and *petN*. The *accD* gene exhibited the greatest average sequence divergence (0.07825), followed by *ycf1* (0.0241), *rps16* (0.0201), and *matK* (0.0194; Fig 5), most of which were located in the LSC region, and the *accD* gene of *G. soja* was highly divergent from those of nine other *Glycine* species (S3 Fig). The highest nucleotide diversity (*Pi*) (0.0916) and total number of mutations (Eta) (119 bp) in comparison with the *G. soja* *accD* gene was observed in *G. cyrtoloba* among the plastomes of the nine *Glycine* species, whereas the lowest were observed in *G. syndetika* (S4 Table). The length of the *accD* gene was 1,299 bp (433 aa) in *G. soja*, *G. max*, and *G. gracilis* and 1527 bp (523 aa) in the seven other *Glycine* species (S3 Fig). Similar differences in gene length within small cpDNA regions have been observed in a variety of other angiosperms [21]. In legume species, both *ycf4* and *accD* exhibit extensive length variation. The expansion of the *accD* gene is partly explained by the presence of numerous tandemly repeated sequences [21]. This *accD* gene encodes a subunit of acetyl-CoA carboxylase, which is related to fatty acid synthesis within the plastid. Previous gene knockout experiments have shown that the function of *accD* is vital, and this gene is expected to be indispensable [101]. However, various studies have identified widespread pseudogenization or absence of *accD* in a variety of relatively distant lineages, including the Ericaceae, Campanulaceae, Geraniaceae, Acoraceae, Poaceae, and Fabaceae [10, 21, 102–106], which implies that deletion or pseudogenization events occur independently.

BOUNDARIES BETWEEN SINGLE-COPY AND IR REGIONS

Variations in the size of angiosperm plastomes are mostly the result of expansion or contraction of the IR regions [79, 107–109]. In the present study, a detailed comparison of the four
Fig 5. Pairwise distance of 76 genes from *Glycine soja* (new and old) and nine other *Glycine* species.

https://doi.org/10.1371/journal.pone.0182281.g005
junctions ($J_{LA}$, $J_{LB}$, $J_{SA}$, and $J_{SB}$) between the two IR regions (IRa and IRb) and the two single-copy regions (LSC and SSC) of the 10 *Glycine* species was performed (Fig 6). Despite the similar lengths of the IR regions of *G. soja* and the other nine *Glycine* species, some expansion and contraction were observed, with the IR regions ranging from 25,432 bp in *G. stenophita* to 25,591 bp in *G. dolichocarpa*. The genes that marked the beginnings and ends of the IR regions were only partially duplicated, including 68 bp of *rpp19* in *G. soja*, *G. max*, and *G. gracilis* and 65 bp of *rpp19* in *G. dolichocarpa*, *G. falcata*, *G. stenophita*, and *G. tomentella*. In *G. canescens*, *G. syndetika*, and *G. cyrtoloba*, this distance was 61 bp in IR region from $J_{LB}$.

Similarly, the hypothetical cp gene *ycf1* was partially duplicated, with 478 bp of this sequence being duplicated in *G. soja*, *G. max*, and *G. gracilis*; 463 bp in *G. falcata*, *G. stenophita*, and *G. tomentella*; and 442 bp in *G. canescens* and *G. cyrtoloba*. $J_{LA}$ was located between *rpl2* and *psbA*, and the
more, the results of the present study suggest that there is no conflict between the complete agreement with the results of Gao and Gao (2016) [37], who reported that based on SSR and plastome data [114, 125]. These results of the present study are in general sets, genome and the 76 shared gene datasets.

Phylogenetic relationships among Glycine species

Plastid genomes have been useful in phylogenetic, evolutionary, and molecular studies. During the last decade, many analyses based on the comparison of plastid protein-coding genes [110, 111] and complete genome sequences [112] have addressed phylogenetic questions at deep nodes and enhanced our understanding of enigmatic evolutionary relationships among angiosperms. The genus Glycine includes 28 species, separated into two subgenera (Soja and Glycine), the former of which includes both cultivated soybean (G. max) and its wild annual progenitor (G. soja), which are distributed in East Asia, including Japan, Korea, China, Russia, and Taiwan. G. max and G. soja are both diploid (2n = 40) and interfertile and are thought to share highly similar genetic variation, although G. soja is much more variable than G. max [25, 113]. Polymorphisms in the cpDNA of G. max and G. soja have been used in numerous studies to assess maternal lineages and cytoplasmic diversity [114–119]. Continued efforts have expanded our ability to differentiate and understand the genomic structure and phylogenetic relationships of Glycine species [28, 120, 121]. The phylogeny and taxonomy of Glycine species in the Soja subgenus have been extensively investigated based on DNA variation, including nucleotide variation in nuclear ribosomal DNA (rDNA), intergenic spacer (ITS) regions [122], cpDNA restriction sites [24, 29], the histone gene H3-D [31], A-199a [123], and cpDNA intergenic spacer regions [25]. However, the complete genome sequence provides more detailed insight [52, 55, 124]. In the present study, the phylogenetic position of G. soja within its genus was established using the complete plastomes (S5 Table) and shared genes of 10 Glycine species and various methods of phylogenetic analysis. Phylogenetic analysis indicated that the complete plastome and the 76 shared genes contained the same phylogenetic signal. In both datasets, G. soja formed a clade with G. max and G. gracilis, with high BI and bootstrap support values (Fig 7, S4 Fig). Moreover, the tree topology confirmed previously reported relationships based on SSR and plastome data [114, 125]. These results of the present study are in general agreement with the results of Gao and Gao (2016) [37], who reported that G. gracilis is intermediate between the two species and is more closely related to G. max than G. soja. Furthermore, the results of the present study suggest that there is no conflict between the complete genome and the 76 shared gene datasets.

Conclusions

In the present study, the complete plastome sequence of G. soja (152,224 bp) was determined. The gene order and structure of the G. soja plastome were found to be highly conserved with the plastomes of other Glycine species. The present study also revealed the distribution and location of repeat sequences and SSRs as well as the sequence divergence among the plastomes and shared genes between G. soja and nine of its congeners. No major structural rearrangement was observed in relation to annual Glycine species. However, in the perennial species, accD was found to be the most divergent gene, while relatively lower identity was observed in some other regions, especially in the rpoC1, atpF, accD, and clpP genes. Furthermore, phylogenetic analyses based on complete plastomes and shared genes yielded trees with the same topology, at least in regard to the placement of G. soja. Thus, the present study provides a
valuable analysis of the complete plastome of *G. soja* and related species, which may facilitate species identification and both biological and phylogenetic studies.

**Supporting information**

**S1 Table.** Primers used for gap closing and sequence verification in *Glycine soja*.

(DOCX)

**S2 Table.** Indel and SNP analysis of the plastid genomes of *Glycine soja* (new and old) and nine other *Glycine* species.

(XLSX)

**S3 Table.** Average pairwise distance of plastid sequences from *Glycine soja* (new and old) and nine other *Glycine* species.

(XLS)

**S4 Table.** Comparison of the nucleotide variability (Pi) and total number of mutations of the *G. soja* accD gene with related species.

(XLSX)
S5 Table. Alignment of complete plastomes from Glycine soja (new and old) and 9 other Glycine species (NEXUS format).

S1 Fig. Visual alignment of plastid genomes from Glycine soja (new) with annual Glycine species (G. soja (old), G. max and G. gracilis). VISTA-based identity plot showing the sequence identity among the ten Glycine species, using G. soja (new) as a reference. Vertical scale indicates the percentage of identity, ranging from 70% to 100%. Horizontal axis indicates the coordinates within the chloroplast genome. Arrows indicate the annotated genes and their transcriptional direction. A thick black line indicates the inverted repeat (IR) regions.

S2 Fig. Sliding window analysis of the complete plastome from Glycine soja (new) with annual Glycine species (G. soja old, G. max and G. gracilis) (Window length: 800 bp, step size: 200 bp). X-axis, position of the midpoint of a window; Y-axis, nucleotide diversity of each window.

S3 Fig. Alignment of accD gene nucleotide sequences among 11 Glycine species plastomes.

S4 Fig. Phylogenetic trees were constructed for ten species from the Glycine genus using different methods, and the Bayesian tree for the whole-genome sequences is shown. The data from the 76 shared genes were analysed with four different methods: joining-joining (NJ), maximum parsimony (MP), maximum likelihood (ML) and Bayesian inference (BI)). The numbers above the branches are the bootstrap values from the NJ, MP, and ML methods and the posterior probabilities of BI. A red dot represents the position of G. soja (KY241814).

Author Contributions
Conceptualization: Sajjad Asaf, Qari Muhammad Imran.
Data Curation: Khdija Al-Hosni.
Formal analysis: Abdul Latif Khan.
Methodology: Abdul Latif Khan.
Resources: Sang-Mo Kang.
Software: Abdul Latif Khan, Qari Muhammad Imran, Sang-Mo Kang, Eun Ju Jeong.
Supervision: In-Jung Lee.
Validation: Ko Eun Lee, In-Jung Lee.
Visualization: Muhammad Aaqil Khan, Sang-Mo Kang.
Writing – original draft: Muhammad Aaqil Khan, Sang-Mo Kang, In-Jung Lee.
Writing – review & editing: Qari Muhammad Imran.

References
1. Neuhaus HE, Emes MJ. NONPHOTOSYNTHETIC METABOLISM IN PLASTIDS. Annu Rev Plant Physiol Plant Mol Biol. 2000; 51:111–40. Epub 2004/03/12. https://doi.org/10.1146/annurev.arplant.51.1.111 PMID: 15012188.
2. Rodríguez-Ezpeleta N, Brinkmann H, Burey SC, Roure B, Burger G, Loffelhardt W, et al. Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. Curr Biol. 2005; 15 (14):1325–30. Epub 2005/07/30. https://doi.org/10.1016/j.cub.2005.06.040 PMID: 16051178.

3. Palmer JD. Plant diversity and evolution: genotypic and phenotypic variation in higher plants: Cabi Publishing; 2005.

4. Henry RJ. Plant diversity and evolution: genotypic and phenotypic variation in higher plants: Cabi Publishing; 2005.

5. Martin G, Baurens FC, Cardi C, Aury JM, D’Hont A. The Complete Chloroplast Genome of Banana (Musa acuminata, Zingiberales): Insight into Plastid Monocotyledon Evolution. Plos One. 2013; 8(6). https://doi.org/10.1371/journal.pone.0067350 PMID: 23840670

6. Lee H-L, Jansen RK, Chumley TW, Kim K-J. Gene relocations within plastid genomes of Jasminum and Menodora (Oleaceae) are due to multiple, overlapping inversions. Mol Biol Evol. 2007; 24 (5):1161–80. https://doi.org/10.1093/molbev/msm229

7. Greiner S, Wang X, Rauwolf U, Silber MV, Mayer K, Meurer J, et al. The complete nucleotide sequences of the five genetically distinct plastid genomes of Oenothera, subsection Oenothera: I. Sequence evolution and plastome evolution. Nucleic Acids Res. 2008; 36(7):2366–78. https://doi.org/10.1093/nar/gkn081 PMID: 18299283

8. Cai Z, Guisinger M, Kim H-G, Ruck E, Blazier JC, McMurtry V, et al. Extensive reorganization of the plastid genome of Tricholium caerulescens are associated with repeats and tRNA genes. J Mol Evol. 2006; 66 (4):350–61. https://doi.org/10.1007/s00239-006-9086-4 PMID: 18330485

9. Wicke S, Schneeweiss GM, dePamphilis CW, Muller KF, Quandt D. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol. 2011; 76(3–5):273–97. https://doi.org/10.1007/s11103-011-9762-4 PMID: 21424877

10. Haberle RC, Fourcade HM, Boore JL, Jansen RK. Extensive rearrangements in the plastid genome of Trachelium caeruleum are associated with repeats and tRNA genes. J Mol Evol. 2008; 66 (4):350–61. https://doi.org/10.1007/s00239-008-9086-4 PMID: 18330485

11. Kim K-J, Choi K-S, Jansen RK. Two chloroplast DNA inversions originated simultaneously during the early evolution of the sunflower family (Asteraceae). Mol Biol Evol. 2005; 22(9):1783–92. https://doi.org/10.1093/molbev/msi174 PMID: 15917497

12. Guisinger MM, Kuehl JV, Boore JL, Jansen RK. Extreme reconfigurations of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and nucleotide substitution rates. Mol Biol Evol. 2013:ms/t257.

13. Weng M-L, Blazier JC, Govindu M, Jansen RK. Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats and nucleotide substitution rates. Mol Biol Evol. 2013:ms/t257.

14. Guisinger MM, Kuehl JV, Boore JL, Jansen RK. Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. Mol Biol Evol. 2011; 28 (1):583–600. https://doi.org/10.1093/molbev/msq229 PMID: 20805190

15. Wojciechowski MF, Lavin M, Sanderson MJ. A phylogeny of legumes (Leguminosae) based on analysis of the plastid matK gene resolves many well-supported subclades within the family. Am J Bot. 2004; 91(11):1846–62. https://doi.org/10.3732/ajb.91.11.1846 PMID: 21652332

16. Doyle JJ, Doyle JL, Ballenger J, Palmer J. The distribution and phylogenetic significance of a 50-kb chloroplast DNA inversion in the flowering plant family Leguminosae. Mol Phylogenet Evol. 1996; 5 (2):429–38. https://doi.org/10.1006/mpev.1996.0038 PMID: 8728401

17. Bruneau A, Doyle JJ, Palmer JD. A chloroplast DNA inversion as a subtribal character in the Phaseoleae (Leguminosae). Syst Bot. 1990:378–86.

18. Doyle JJ, Doyle JL, Palmer JD. Multiple independent losses of two genes and one intron from legume chloroplast genomes. Syst Bot. 1995:272–94.

19. Jansen RK, Wojciechowski MF, Sanniyasi E, Lee S-B, Daniell H. Complete plastid genome sequence of the chickpea (Cicer arietinum) and the phylogenetic distribution of rps12 and clpP intron losses among legumes (Leguminosae). Mol Phylogenet Evol. 2008; 48(3):1204–17. https://doi.org/10.1016/j.ympev.2008.06.013 PMID: 18638561

20. Ganitt JS, Baldauf SL, Calie PJ, Weeden NF, Palmer JD. Transfer of rpl22 to the nucleus greatly preceded its loss from the chloroplast and involved the gain of an intron. The EMBO Journal. 1991; 10 (10):3073. PMID: 1915281

21. Magee AM, Aspinall S, Rice DW, Cusack BP, Sémon M, Perry AS, et al. Localized hypermutation and associated gene losses in legume chloroplast genomes. Genome Res. 2010; 20(12):1700–10. https://doi.org/10.1101/gr.111955.110 PMID: 20978141
22. Singh R, Hymowitz T. The genomic relationships among six wild perennial species of the genus Glycine subgenus Glycine Willd. Theor Appl Genet. 1985; 71(2):221–30. https://doi.org/10.1007/BF00252059 PMID: 24247386

23. Ratnaparkhe M, Singh R, Doyle J. Glycine. Wild Crop Relatives: Genomic and Breeding Resources: Springer; 2011. p. 83–116.

24. Doyle JJ, Doyle JL, Brown A. A chloroplast-DNA phylogeny of the wild perennial relatives of soybean (Glycine subgenus Glycine): congruence with morphological and crossing groups. Evolution. 1990;37:1–89. https://doi.org/10.2307/24247386

25. Sakai M, Kanazawa A, Fujii A, Thseng F, 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(1):29–54.

26. Xu D, Abe J, Gai J, Shimamoto Y. Diversity of chloroplast DNA SSRs in wild and cultivated soybeans: evidence for multiple origins of cultivated soybean. Theor Appl Genet. 2002; 105(5):645–53. https://doi.org/10.1007/s00122-002-0972-7 PMID: 12582476

27. Xu D, Abe J, Kanazawa A, Gai J, Shimamoto Y. Identification of sequence variations by PCR-RFLP and its application to the evaluation of cpDNA diversity in wild and cultivated soybeans. Theor Appl Genet. 2001; 102(5):683–8.

28. Doyle JJ, Doyle JL, Brown A. Analysis of a polyploid complex in Glycine with chloroplast and nuclear DNA. Australian Systematic Botany. 1990; 3(1):125–36.

29. DOYLE JJ, DOYLE JL, RAUSCHER JT, Brown A. Evolution of the perennial soybean polyploid complex (Glycine subgenus Glycine): a study of contrasts. Biol J Linn Soc. 2004; 82(4):583–97.

30. Doyle JJ, Doyle JL, Brown A. Incongruence in the diploid B-genome species complex of Glycine (Leguminosae) revisited: histone H3-D alleles versus chloroplast haplotypes. Mol Biol Evol. 1999; 16(3):354–62. PMID: 10331262

31. Doyle JJ, Doyle JL, Rauscher JT, Brown A. Diploid and polyploid reticulate evolution throughout the history of the perennial soybeans (Glycine subgenus Glycine). New Phytol. 2004; 161(1):121–32.

32. Shinozaki K, Ohme M, Tanaka M, Wakaegi T, Hayashida N, Matsuyoshi T, et al. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. The EMBO journal. 1986; 5(9):2043. PMID: 16453699

33. Daniell H, Lin C-S, Yu M, Chang W-J. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biol. 2016; 17(1):134. https://doi.org/10.1186/s13059-016-1004-2 PMID: 27339192

34. Saski C, Lee S-B, Daniell H, Wood TC, Tomkins J, Kim H-G, et al. Complete chloroplast genome sequence of Glycine max and comparative analyses with other legume genomes. Plant Mol Biol. 2005; 59(2):309–22. https://doi.org/10.1007/s11103-005-0882-0 PMID: 16247559

35. 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(10):1651–65. https://doi.org/10.3732/ajb.1400121 PMID: 25326613

36. Gao C-W, Gao L-Z. The complete chloroplast genome sequence of wild soybean, Glycine soja. Conservation Genetics Resources. 2016:1–3. https://doi.org/10.1071/s12686-016-0659-z

37. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, et al. The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nat Genet. 2011; 43(5):476–+. https://doi.org/10.1038/ng.807 PMID: 21478890

38. Wyman SK, Jansen RK, Boore JL. Automatic annotation of organellar genomes with DOGMA. Bioinformatics. 2004; 20(17):3252–5. https://doi.org/10.1093/bioinformatics/bth352 PMID: 15180927

39. Schattner P, Brooks AN, Lowe TM. The IRNASA-SE, tRNAscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. 2005; 33:W686–W9. https://doi.org/10.1093/nar/gkj366 PMID: 15980563

40. Lohe M, Drehsel O, Bock R. OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Curr Genet. 2007; 52(5-6):267–74. https://doi.org/10.1007/s00294-007-0161-y PMID: 17957369

41. Kumar S, Nei M, Dudley J, Tamura K. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform. 2008; 9(4):299–306. Epub 2008/04/18. https://doi.org/10.1093/bib/bbn017 PMID: 18417537;

42. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 2004; 32:W273–W9. https://doi.org/10.1093/nar/gkh458 PMID: 15215394
44. 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(22):4633–42. https://doi.org/10.1093/nar.29.22.4633 PMID: 11713313

45. Kraemer L, Beszteri B, Gäbler-Schwarz S, Held C, Leese F, Mayer C, et al. STAMP: Extensions to the STADEN sequence analysis package for high throughput interactive microsatellite marker design. Bmc Bioinformatics. 2009; 10(1):41. https://doi.org/10.1186/1471-2105-10-41 PMID: 19183437

46. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999; 27(2):573–80. Epub 1998/12/24. PMID: 9862982

47. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013; 30(4):772–80. https://doi.org/10.1093/molbev/msr010 PMID: 23329690

48. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980; 16(2):111–20. Epub 1980/12/01. PMID: 7463489.

49. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009; 25(11):1451–2. https://doi.org/10.1093/bioinformatics/btp187 PMID: 19346325

50. Morton BR. Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages. J Mol Evol. 1998; 46(4):449–59. https://doi.org/10.1007/Pl00006325 PMID: 9541540

51. Nie XJ, Lv SZ, Zhang YX, Du XH, Wang L, Biradar SS, et al. Chloroplast genome sequence of the major invasive species, Crofton Weed (Ageratina adenophora). Plos One. 2012; 7(5). https://doi.org/10.1371/journal.pone.0036869 PMID: 22906302

52. Oliver MJ, Murdock AG, Mishler BD, Kuehl JV, Boore JL, Mandoli DF, et al. Chloroplast genome sequence of the moss Tortula ruralis: gene content, polymorphism, and structural arrangement relative to other green plant chloroplast genomes. Bmc Genomics. 2010; 11. https://doi.org/10.1186/1471-2164-11-143 PMID: 20187961

53. Wicke S, Schneeweiß GM, dePamphilis CW, Müller KF, Quandt D. The evolution of the plastid chromosome in land plants: gene content, gene order, and functional conservation. Plant Mol Biol. 2011; 76(3–5):251–61. https://doi.org/10.1007/s11103-010-9762-4 PMID: 21424877

54. Jansen RK, Cai Z, Raubeson LA, Daniell H, Leebens-Mack J, Müller KF, et al. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proceedings of the National Academy of Sciences. 2007; 104(49):19369–74.
66. Nakkaew A, Chotigeat W, Eksomtrama T, Phongdara A. Cloning and expression of a plastid-encoded subunit, beta-carboxyltransferase gene (accD) and a nuclear-encoded subunit, biotin carboxylase of acetyl-CoA carboxylase from oil palm (Elaeis guineensis Jacq.). Plant Sci. 2008; 175(4):497–504.

67. Sherman-Broyles S, Bombarely A, Grimwood J, Schmutz J, Doyle J. Complete plastome sequences from Glycine syntetika and six additional perennial wild relatives of soybean. G3 (Bethesda). 2014; 4. https://doi.org/10.1534/g3.114.012690 PMID: 25155272

68. Sato S, Nakamura Y, Kaneko T, Asamizu E, Tabata S. Complete structure of the chloroplast genome of Arabidopsis thaliana. DNA Res. 1999; 6(5):283–90. Epub 1999/11/26. PMID: 10574454.

69. Xu Q, Xiong G, Li P, He F, Huang Y, Wang K. Analysis of complete nucleotide sequences of 12 Gossypium chloroplast genomes: origin and evolution of allotetraploids. Plos One. 2012; 7. https://doi.org/10.1371/journal.pone.0037128 PMID: 22876273

70. Huang H, Shi C, Liu Y, Mao SY, Gao LZ. Thirteen Camellia chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. Bmc Evol Biol. 2014; 14. https://doi.org/10.1186/1471-2148-14-151 PMID: 25001059

71. Adachi Y, Kuroda H, Yukawa Y, Sugirua M. Translation of partially overlapping psbD-psbC mRNAs in chloroplasts: the role of 5'-processing and translational coupling. Nucleic Acids Res. 2012; 40(7):3152–8. https://doi.org/10.1093/nar/gkr1185 PMID: 22156163

72. Yukawa M, Sugirua M. Termination codon-dependent translation of partially overlapping ndhC-ndhK transcripts in chloroplasts. Proc Natl Acad Sci U S A. 2008; 105(49):19550–4. Epub 2008/11/27. https://doi.org/10.1073/pnas.0809240105 PMID: 19033452;

73. Cavalier-Smith T. Chloroplast evolution: Secondary symbiogenesis and multiple losses. Curr Biol. 2002; 12(2):R62–R4. https://doi.org/10.1016/S0960-9822(01)00675-3 PMID: 11818081

74. Yi X, Gao L, Wang B, Su YJ, Wang T. The Complete Chloroplast Genome Sequence of Cephalotaxus oliveri (Cephalotaxaceae): Evolutionary Comparison of Cephalotaxus Chloroplast DNA and Insights into the Loss of Inverted Repeat Copies in Gymnosperms. Genome Biol Evol. 2012; 5(4):688–98. https://doi.org/10.1093/gbe/evt042 PMID: 23538991

75. Asano T, Tsudzuki T, Takahashi S, Shimada H, Kadowaki K. Complete nucleotide sequence of the sugarcan (Saccharum officinarum) chloroplast genome: A comparative analysis of four monocot chloroplast genomes. DNA Res. 2004; 11(2):93–9. https://doi.org/10.1093/dnares/11.2.93 PMID: 15449542

76. Timme RE, Kuehl JV, Boore JL, Jansen RK. A comparative analysis of the Lactuca and Helianthus (Asteraceae) plastid genomes: Identification of divergent regions and categorization of shared repeats. Am J Bot. 2007; 94(3):302–12. https://doi.org/10.3732/ajb.94.3.302 PMID: 21636403

77. Gao L, Yi X, Yang YX, Su YJ, Wang T. Complete chloroplast genome sequence of a tree fern Alsophila spinulosa: insights into evolutionary changes in fern chloroplast genomes. Bmc Evol Biol. 2009; 9. https://doi.org/10.1186/1471-2148-9-130 PMID: 19519899

78. Rose O, Falush D. A threshold size for microsatellite expansion. Mol Biol Evol. 1998; 15(5):613–5. PMID: 9580993

79. Raubeson LA, Peery R, Chumley TW, Dziubek C, Fourcade HM, Boore JL, et al. Comparative chloroplast genomics: analyses including new sequences from the angiosperms Nuphar advena and Ranunculus manchuricus. Bmc Genomics. 2007; 8. https://doi.org/10.1186/1471-2164-8-174 PMID: 17573971

80. Huotari T, Korpelainen H. Complete chloroplast genome sequence of Elodea canadensis and comparative analyses with other monocot plastid genomes. Gene. 2012; 508(1):96–105. https://doi.org/10.1016/j.gene.2012.07.020 PMID: 22841789

81. Powell W, Morgante M, Mcdevitt R, Vendramin GG, Rafalski JA. Polymorphic Simple Sequence Repeat Regions in Chloroplast Genomes—Applications to the Population-Genetics of Pines. P Natl Acad Sci USA. 1999; 92(17):7759–63. https://doi.org/10.1073/pnas.92.17.7759

82. Pauwels M, Vekemans X, Gode C, Frerot V, Castric V, Saumitou-Laprade P. Nuclear and chloroplast DNA phylogeny reveals vicariance among European populations of the model species for the study of metal tolerance, Arabidopsis halleri (Brassicaceae). New Phytol. 2012; 193(4):916–28. https://doi.org/10.1111/nph.12030 PMID: 23228532

83. Powell W, Morgante M, Andre C, Mcnicol JW, Machray GC, Doyle JJ, et al. Hypervariable Microsatellites Provide a General Source of Polymorphic DNA Markers for the Chloroplast Genome. Curr Biol. 1995; 5(9):1023–9. https://doi.org/10.1016/S0960-9822(95)00206-5 PMID: 8542278
85. Li XW, Gao HH, Wang YT, Song JY, Henry R, Wu HZ, et al. Complete chloroplast genome sequence of Magnolia grandiflora and comparative analysis with related species. Sci China Life Sci. 2013; 56(2):189–98. https://doi.org/10.1007/s11427-012-4430-8 PMID: 23329156

86. Kuang DY, Wu H, Wang YL, Gao LM, Zhang SZ, Lu L. Complete chloroplast genome sequence of Magnolia kwangsiensis (Magnoliaceae): implication for DNA barcoding and population genetics. Genome. 2011; 54(8):663–73. https://doi.org/10.1139/G11-026 PMID: 21793699

87. Palmer JD, Osorio B, Thompson WF. Evolutionary significance of inversions in legume chloroplast DNAs. Curr Genet. 1988; 14. https://doi.org/10.1007/bf00405856

88. Kazakoff SH, Imelfort M, Edwards D, Koehorst J, Biswas B, Batley J. Capturing the biofuel wellhead and powerhouse: the chloroplast and mitochondrial genomes of the leguminous feedstock tree Pongamia pinnata. Plos One. 2012; 7. https://doi.org/10.1371/journal.pone.0051687 PMID: 23272141

89. Tangphatsornruang S, Sangsrakru D, Chanprasert J, Uthaipaisanwong P, Yoocha T, Jomchai N, et al. The Chloroplast Genome Sequence of Mungbean (Vigna radiata) Determined by High-throughput Pyrosequencing: Structural Organization and Phylogenetic Relationships. DNA Res. 2010; 17(1):11–22. https://doi.org/10.1093/dnares/dsp025 PMID: 20007682

90. Martin GE, Rousseau-Gueutin M, Cordonnier S, Lima O, Naquin D, et al. The first complete chloroplast genome of the Genistoid legume Lupinus luteus: evidence for a novel major lineage-specific rearrangement and new insights regarding plastome evolution in the legume family. Ann Bot-London. 2014; 113(7):1197–210. https://doi.org/10.1093/aob/mcu050 PMID: 24769537

91. Perry AS, Brennan S, Murphy DJ, Kavanagh TA, Wolfe KH. Evolutionary re-organisation of a large operon in adzuki bean chloroplast DNA caused by inverted repeat movement. DNA Res. 2002; 9(5):157–62. https://doi.org/10.1093/dnares/9.5.157 PMID: 12465715

92. Millien RS, Olmstead RG, Adams KL, Palmer JD, Lao NT, Heggie L. Many parallel losses of infA from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. Plant Cell. 2001; 13. https://doi.org/10.1105/tpc.13.3.645

93. Sabir J, Schwarz E, Ellison N, Zhang J, Baeshen NA, Mutwakil M, et al. Evolutionary and biotechnology implications of plastid genome variation in the inverted-repeat-lacking clade of legumes. Plant Biotechnol J. 2000; 101(5):724–32. https://doi.org/10.1093/aob/mcu050 PMID: 24769537

94. Kumar S, Hahn FM, McMahan CM, Cornish K, Whalen MC. Comparative analysis of the complete sequence of the plastid genome of Parthenium argentatum and identification of DNA barcodes to differentiate Parthenium species and lines. Bmc Plant Biol. 2009; 9. https://doi.org/10.1186/1471-2229-9-131 PMID: 19917140

95. Xu DH, Abe J, Sakai M, Kanazawa A, Shimamoto Y. Sequence variation of non-coding regions of chloroplast DNA of soybean and related wild species and its implications for the evolution of different chloroplast haplotypes. Theor Appl Genet. 2000; 101(5):724–32. https://doi.org/10.1007/s001220051537

96. Kanazawa A, Tozuka A, Shimamoto Y. Sequence variation of chloroplast DNA that involves EcoRI and CiaI restriction site polymorphisms in soybean. Genes Genet Syst. 1998; 73(2):111–9. Epub 1998/08/27. PMID: 9718676.

97. Spielmann A, Stutz E. Nucleotide sequence of soybean chloroplast DNA regions which contain the psb A and trn H genes and cover the ends of the large single copy region and one end of the inverted repeats. Nucleic Acids Res. 1983; 11(20):7157–67. PMID: 6314279

98. Skuza L, Filip E, Szućko I. Use of organelle markers to study genetic diversity in soybean. 2013.

99. Small RL, Ryburn JA, Cronn RC, Seelanan T, Wendel JF. The tortoise and the hare: choosing between noncoding plastome and nuclear Adh sequences for phylogeny reconstruction in a recently diverged plant group. Am J Bot. 1998; 85(9):1301–15. PMID: 21685016

100. Demesure B, Comps B, Petit RJ. Chloroplast DNA phylogeography of the common beech (Fagus sylvatica L.) in Europe. Evolution. 1996; 50(6):2515–20. https://doi.org/10.1111/j.1558-5646.1996.tb03636.x PMID: 28565658

101. Kode V, Mudd EA, Iamtham S, Day A. The tobacco plastid accD gene is essential and is required for leaf development. Plant J. 2005; 44(2):237–44. https://doi.org/10.1111/j.1365-313X.2005.02533.x PMID: 16212603

102. Goremykin VV, Holland B, Hirsch-Ernst KI, Hellwig FH. Analysis of Acorus calamus chloroplast genome and its phylogenetic implications. Mol Biol Evol. 2005; 22(9):1813–22. https://doi.org/10.1093/molbev/msi173 PMID: 15930156

103. Guisinger MM, Kuehl JV, Boore JL, Jansen RK. Genome-wide analyses of Geraniaceae plastid DNA reveal unprecedented patterns of increased nucleotide substitutions. Proceedings of the National Academy of Sciences. 2008; 105(47):18424–9.
104. Fajardo D, Senalik D, Ames M, Zhu H, Steffan SA, Harbut R, et al. Complete plastid genome sequence of Vaccinium macrocarpon: structure, gene content, and rearrangements revealed by next generation sequencing. Tree Genet Genomes. 2013; 9(2):489–98.

105. Harris ME, Meyer G, Vandergon T, Vandergon VO. Loss of the acetyl-CoA carboxylase (accD) gene in Poales. Plant Mol Biol Rep. 2013; 31(1):21–31.

106. Martínez-Alberola F, del Campo EM, Lázaro-Gimeno D, Mezquita-Claramonte S, Molinas A, Mateu-Andrés I, et al. Balanced gene losses, duplications and intensive rearrangements led to an unusual regularly sized genome in Arbutus unedo chloroplasts. Plos One. 2013; 8(11):e79685. https://doi.org/10.1371/journal.pone.0079685 PMID: 24260278

107. Yang M, Zhang X, Liu G, Yin Y, Chen K, Yun Q. The complete chloroplast genome sequence of date palm (Phoenix dactylifera L.). Plos One. 2010; 5. https://doi.org/10.1371/journal.pone.0012762 PMID: 20856810

108. Wang RJ, Cheng CL, Chang CC, Wu CL, Su TM, Chaw SM. Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. Bmc Evol Biol. 2008; 8. https://doi.org/10.1186/1471-2148-8-36 PMID: 18237435

109. Yang Y, Dang YY, Li Q, Li JJ, Li WX, Wang YT. Complete Chloroplast Genome Sequence of Poisonous and Medicinal Plant Datura stramonium: Organizations and Implications for Genetic Engineering. Plos One. 2014; 9(11). https://doi.org/10.1371/journal.pone.0079656 PMID: 25365514

110. Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. Proceedings of the National Academy of Sciences. 2010; 107(10):4623–8. https://doi.org/10.1073/pnas.0907801107 PMID: 20176954

111. Goremykin VV, Hirsch-Ernest KI, Wolff S, Hellwig FH. The chloroplast genome of Nymphaea alba: Whole-genome analyses and the problem of identifying the most basal angiosperm. Mol Biol Evol. 2004; 21(7):1445–54. https://doi.org/10.1093/molbev/msh147 PMID: 15084683

112. Hohmann N, Schmickl R, Chiang T-Y, Lúčanova M, Kolář F, Marhold K, et al. Taming the wild: resolving the gene pools of non-model Arabidopsis lineages. Bmc Evol Biol. 2014; 14(1):1–21. https://doi.org/10.1186/s12862-014-0224-x PMID: 25344686

113. Singh RJ, Hymowitz T. The genomic relationship between Glycine max (L.) Merr. and G. soja Sieb. and Zucc. as revealed by pachytene chromosome analysis. Theor Appl Genet. 1988; 76(5):705–11. Epub 1988/11/01. https://doi.org/10.1007/BF00303516 PMID: 24232348

114. Shoemaker R, Hatfield P, Palmer R, Athery A. Chloroplast DNA variation in the genus Glycine subgenus Soja. J Hered. 1986; 77(1):26–30.

115. Close P, Shoemaker R, Keim P. Distribution of restriction site polymorphism within the chloroplast genome of the genus Glycine, subgenus Soja. Theor Appl Genet. 1989; 77(6):768–76. https://doi.org/10.1007/BF00268325 PMID: 24232890

116. Hirata T, Abe J, Shimamoto Y. RFLPs of chloroplast and mitochondrial genomes in summer and autumn maturing cultivar groups of soybean in Kyushu district of Japan. Soybean genetics newsletter (USA). 1996.

117. Lee D, Caha C, Specht J, Graef G. Chloroplast DNA evidence for non-random selection of females in an outcrossed population of soybeans [Glycine max (L.)]. Theor Appl Genet. 1992; 85(2–3):261–8. https://doi.org/10.1007/BF00222868 PMID: 24197313

118. Shimamoto Y, Hasegawa A, Abe J, Ohara M, Mikami T. Glycine soja germplasm in Japan: isozyme and chloroplast DNA variation. Soybean genetics newsletter-US Department of Agriculture, Agricultural Research Service (USA). 1992.

119. Abe J, Hasegawa A, Fukushi H, Mikami T, Ohara M, Shimamoto Y. Introgression between wild and cultivated soybeans of Japan revealed by RFLP analysis for chloroplast DNA. Economic Botany. 1999; 53(3):285–91.

120. Close P, Shoemaker R, Keim P. Distribution of restriction site polymorphism within the chloroplast genome of the genus Glycine, subgenus Soja. TAG Theoretical and Applied Genetics. 1989; 77 (6):768–76. https://doi.org/10.1007/BF00268325 PMID: 24232890

121. Kollipara KP, Singh RJ, Hymowitz T. Phylogenetic and genomic relationships in the genus Glycine Wild. based on sequences from the ITS region of nuclear rDNA. Genome. 1997; 40(1):57–68. Epub 1997/02/01. PMID: 9061914.

122. Zhu T, Schupp JM, Olphant A, Keim P. Hypomethylated Sequences—Characterization of the Duplicated Soybean Genome. Mol Genet. 1994; 244(6):638–45. PMID: 7969033
124. Wambugu P, Brozynska M, Furtado A, Waters D, Henry R. Relationships of wild and domesticated rices (Oryza AA genome species) based upon whole chloroplast genome sequences. Sci Rep. 2015; 5. https://doi.org/10.1038/srep13957 PMID: 26355750

125. Gao C-W, Gao L-Z. The complete chloroplast genome sequence of semi-wild soybean, Glycine gracilis (Fabales: Fabaceae). Conservation Genetics Resources. 2017;1–3.