

Comparative analysis of the complete chloroplast genome of jack bean (*Canavalia ensiformis*, Fabaceae) revealed mutational hotspots and phylogenetic relationships

Isna Arofatur Nikmah

isna.nikmah@idu.ac.id

Department of Biology, Faculty of Military Mathematics and Natural Sciences, The Republic Indonesia Defense University. Komplek Indonesia Peace and Security Center (IPSC) Sentul. Bogor, Indonesia 16810
<https://orcid.org/0000-0002-6994-8583>

Wendy A. Mustaqim

Program Studi Biologi, Fakultas Sains dan Teknologi, Universitas Samudra, Jl. Prof. Dr. Syarif Thayeb, Meurandeh, Langsa Lama, Langsa, Aceh 24416, Indonesia <https://orcid.org/0000-0001-9902-830X>

Wisnu Handoyo Ardi

Research Center for Biosystematics and Evolution, Research Organization for Life Sciences and Environment, National Research and Innovation Agency (BRIN), Bogor Indonesia 16911 <https://orcid.org/0000-0001-7388-7608>

Rugayah

Research Center for Biosystematics and Evolution, Research Organization for Life Sciences and Environment, National Research and Innovation Agency (BRIN), Bogor Indonesia 16911 <https://orcid.org/0000-0003-1431-1531>

Rizky Dwi Satrio

rizky.satrio@idu.ac.id

Department of Biology, Faculty of Military Mathematics and Natural Sciences, The Republic Indonesia Defense University. Komplek Indonesia Peace and Security Center (IPSC) Sentul. Bogor, Indonesia 16810
<https://orcid.org/0000-0003-1716-2076>

Research Article

Keywords: Fabaceae, Genome, Plastid, Underutilized crop

Posted Date: March 3rd, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-6125303/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

In recent years, the study of chloroplast genomes has gained increasing attention due to its significance in understanding plant evolution, phylogenetics, and genetic engineering. *Canavalia ensiformis*, commonly known as jack bean, is a leguminous plant of agricultural and ecological importance, with its chloroplast genome offering potential insights into adaptive mechanisms, species divergence, and biotechnological applications. This study focuses on the structure, organization, and gene content of the *C. ensiformis* chloroplast genome and its comparative analysis with other members of Phaseoleae (Fabaceae). We assembled and annotated the complete chloroplast genome of *C. ensiformis* using a combination of Illumina short-reads and Oxford Nanopore long-reads available on GenBank. The *C. ensiformis* chloroplast genome was organized as a single circular structure of 157,809 bp in length and divided into four main regions: a large single-copy (77,529 bp), a small single-copy (18,934 bp), and a pair of inverted repeat regions (30,673 bp). The genome encoded 95 protein-coding genes, 8 rRNA, and 37 tRNA genes. To explore evolutionary relationships, we conducted a comparative analysis of plastid genomes across Phaseoleae species. Structural variations, gene rearrangements, and divergence in non-coding regions were assessed, revealing conserved synteny and lineage-specific differences within the tribe. The phylogenomic analysis based on plastid genome sequences further clarified the evolutionary placement of *C. ensiformis* within Phaseoleae. The results contribute to a broader understanding of plastid genome evolution in legumes, one of the most species-rich plant families, and provide valuable genomic resources for future research in molecular phylogenetics and crop improvement.

Introduction

Canavalia ensiformis, commonly known as the jack bean, is a robust leguminous plant belonging to the Fabaceae family (Patel et al., 2016). Morphologically, it is characterized by its vigorous climbing or trailing growth habit, trifoliate leaves with ovate to elliptical leaflets, and purplish-pink flowers borne in axillary racemes (Sauer, 1964; Rosenthal, 1971; Sahai, 2009). The plant produces elongated, cylindrical pods containing several large, smooth seeds, typically white or light brown in color (Rodríguez–Quiñones et al., no date; Eke, Asoegwu and Nwandikom, 2007). Native to tropical and subtropical regions, *C. ensiformis* thrives in poor soils and harsh environmental conditions, aided by its symbiotic nitrogen-fixing capabilities (Ekanayake, Jansz and Nair, 2000). It is widely recognized for its versatility in agricultural and ecological applications (Ponganan et al., 2024). As a cover crop, it improves soil fertility, suppresses weeds, and prevents erosion, while its high-protein seeds serve as livestock feed and, with proper processing, a food source for humans (Savita, 2023). Additionally, *C. ensiformis* has garnered attention for its bioactive compounds, including canavanine, which exhibits antimicrobial and pesticidal properties (Downum, Rosenthal and Cohen, 1983). This multifunctional species plays a crucial role in sustainable agriculture and holds promise for further applications in food security and environmental management.

In addition, the study of phylogenetic relationships within the *Canavalia* genus, including *C. ensiformis*, has received increasing attention due to the ecological and agricultural importance of these species. Researchers aim to unravel the evolutionary relationships among *Canavalia* species to better understand their adaptation, genetic diversity, and functional traits (Snak et al., 2016). These studies often involve morphological (Sauer, 1964), biochemical, and molecular approaches (Snak et al., 2016; Suvarna and Sharma, 2020). However, the overlapping morphological characteristics among *Canavalia* species, such as similar flower structures and pod

morphology, have posed challenges in distinguishing them solely based on phenotypic traits (Sauer, 1964; Sahai, 2009). Consequently, molecular phylogenetics has become the primary method for studying relationships within this genus.

Plastomes, or chloroplast genomes, offer significant advantages for studying phylogenetic relationships, particularly in species *Canavalia ensiformis*. These genomes are highly conserved in structure (Guo et al., 2017) and gene content (Wicke et al., 2011) across plant taxa, making them ideal for comparative analyses. Their uniparental inheritance, typically maternal (Schneider, 2023) minimizes recombination and provides a clear evolutionary signal (0000-0003-0136-735X et al., no date; Chen et al., 2023; Dong et al., 2023). Additionally, plastomes contain both conserved coding regions, such as *rbcL* and *matK*, and highly variable non-coding regions (Gargano et al., 2012) offering flexibility to resolve relationships at various taxonomic levels (Song et al., 2020). The high copy number of plastomes in plant cells makes them easy to isolate and sequence, even from challenging samples ((Huang et al., 2014). In *C. ensiformis* and related species, plastome-based phylogenetic studies can reveal evolutionary relationships, assess genetic diversity, and clarify speciation events, providing critical insights for taxonomy, conservation, and agricultural applications.

The plastid genome of *Canavalia ensiformis* has been studied, however there is a lack of fully assembled and annotated complete genome sequences in some databases. In many cases, available sequences are partial or fragmented, which limits comprehensive comparative analysis. Having a fully sequenced plastid genome would enable more accurate and in-depth phylogenetic studies and functional investigations into plastid-specific processes like photosynthesis and metabolic pathways. Therefore, this study aims to sequence the plastid genomes of *Canavalia ensiformis*; To perform comparative analysis with other species of legumes or related taxa; To characterize unique structural or functional features in the genomes of *Canavalia*.

Materials and Methods

Fresh leaf sampling and DNA sequencing

Jack bean plants were selected and cultivated under natural environmental conditions; fresh leaves from one bud and two additional leaves were collected and promptly preserved in liquid nitrogen. Upon returning the samples to the laboratory, they were pulverised and preserved at – 80°C before experimental analysis. Genomic DNA was extracted utilising the cetrimonium bromide method. Following quality control, the genomic DNA was analysed using the NextSeq 2000 platform (Illumina, Inc.) and the MinION platform (Oxford Nanopore Technologies).

Genome assembly and annotation

This study utilised fastp software (<https://github.com/OpenGene/fastp>) to filter the raw data. The criteria for filtering were as follows: (1) truncate the sequencing linker and primer sequences from the reads; (2) exclude reads with an average quality value below Q5; and (3) eliminate reads containing more than 5 N's. The original sequencing data were compared with a self-constructed chloroplast genome database using Bowtie2's very-sensitive-local pattern ratios, and the aligned sequence was designated as the chloroplast genome sequence for the project sample. Assembly was conducted utilising Spades v3.10.1 software

(<http://cab.spbu.ru/software/spades/>) (Tillich et al. 2017); the chloroplast genome was annotated through two methods to enhance annotation precision. Initially, Prodigal v2.6.3 (<https://www.github.com/hyattpd/Prodigal>) was employed to annotate chloroplast protein-coding genes (PCGs), Hmmer v3.1b2 (<http://www.hmmer.org/>) software was utilised for rRNA prediction, and Aragorn v1.2.38 (<http://130.235.244.92/ARAGORN/>) was applied for tRNA prediction. The second annotation result was derived by extracting the gene sequence of related taxa published on the National Centre for Biotechnology Information, part of the US National Institutes of Health, followed by the use of Blast v2.6 (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) to compare the assembled sequence. The two annotation results were manually verified for various genes, erroneous and redundant annotations were eliminated, and the multi-exonic boundary was established to finalise the annotation. The annotated sequence was submitted to the NCBI database under accession number PV174432.

Comparison and marker identification

The IRscope v2.01 software was employed to examine the limits of inverted repeats and their contraction and expansion in 60 chloroplast genomes of Fabaceae species. DnaSP software version 6.0 (Rozas et al. 2017) was employed to ascertain the LSC, SSC, and IR regions, in addition to calculating the nucleotide diversity (π) for the Fabaceae species. A sliding window analysis employing a 600-bp window and 200-bp increments was utilised to plot π values.

Relative codon usage analysis

The RSCU analysis of coding genes in the chloroplast genome was performed utilising the Genepioneer cloud service (cloud.genepioneer.com) with default parameters.

Repeat identifications

The forward, reverse, palindromic, and complementary repeat sequences were identified utilising REPuter (<https://bibiserv.cebitec.unibielefeld.dereputer/>) with the parameters of a Hamming distance of 3, a minimum repeat size of 30 bp, and an e-value of less than $1e-05$. Furthermore, simple sequence repeats within the assembled chloroplast genomes were detected utilising the MISA website (<https://webblast.ipk-gatersleben.de/misa/>), with the minimum thresholds for mono-, di-, tri-, tetra-, penta-, and hexanucleotides established at 8, 5, 3, 3, 3, and 3, respectively. To account for the potential presence of dispersed repeats across various chromosomes, the chromosomes were artificially concatenated to identify these repeats, after which the results were segmented.

Phylogenetic analysis

To elucidate the phylogenetic placement of *C. duntsa* within the genus *Camellia*, phylogenetic analyses were conducted on 60 chloroplast genomes of Phaseoleae (Fabaceae) taxa, utilising *Pisum sativum* as an outgroup. The chloroplast sequences were aligned utilising MAFFT with the default settings. Subsequently, RAxML software version 7.2.8 (version 8.2.10, <https://cme.h-its.org/exelixis/software.html>) was employed to select a GTRGAMMA model with a bootstrap of 1000 for constructing the maximum likelihood evolutionary tree.

Bayesian analysis was conducted utilising MrBayes v3.2.7 (v3.2.7a, <http://nbisweden.github.io/MrBayes/>), employing Markov Chain Monte Carlo iterative processes for 1 million generations, with sampling occurring every 100 generations.

Results

Plastid Genome of *Canavalia ensiformis*

The complete plastid genome of *Canavalia ensiformis* is a circular DNA molecule with a total length of 157,809 bp, displaying a typical quadripartite structure consisting of a large single-copy (LSC) region (77,529 bp), a small single-copy (SSC) region (43,981 bp), and two inverted repeat (IR) regions (each 37,603 bp) (**Fig. 1**). The GC content of the genome is approximately 35%, which is within the expected range for leguminous plants. A total of 136 genes were annotated, including 89 protein-coding genes, 8 ribosomal RNA (rRNA) genes, and 39 transfer RNA (tRNA) genes. The genome encodes essential components for photosynthesis, including genes related to Photosystem I (*psaA*, *psaB*), Photosystem II (*psbA*, *psbB*), cytochrome b6f (*petA*, *petB*), and ATP synthesis (*atpA*, *atpB*), which are critical for the plant's energy metabolism. Additionally, genes involved in carbon fixation, such as *rbcL*, encoding the large subunit of Rubisco, are present and play a fundamental role in photosynthetic CO₂ assimilation. The genome also contains genes associated with transcription and translation, including RNA polymerase subunits (*rpoA*, *rpoB*, *rpoC1*, *rpoC2*) and ribosomal proteins (*rps*, *rpl*), which are vital for plastid gene expression and maintenance. The presence of genes encoding NADH dehydrogenase (*ndhA–ndhK*) suggests an alternative electron transport mechanism supporting photosynthesis under stress conditions.

A notable feature of the *C. ensiformis* plastid genome is the presence of structural and functional elements indicative of evolutionary conservation and plastome stability. The *clpP* gene, which encodes a proteolytic subunit involved in plastid protein homeostasis, contains introns that contribute to post-transcriptional regulation. Likewise, the *matK* gene, often used in phylogenetic studies due to its high rate of nucleotide substitution, is present and suggests potential for evolutionary comparisons within Fabaceae. The inverted repeat (IR) regions harbor rRNA genes (*rrn16*, *rrn23*, *rrn4.5*, *rrn5*) and contribute to genome stability by preventing extensive recombination events. Notably, the plastome contains hypothetical open reading frames (*ycf* genes), whose functions remain unclear but may be involved in plastid biogenesis.

The plastid genome of *Canavalia ensiformis* encodes a diverse array of genes involved in essential biological functions, including photosynthesis, transcription, and translation (**Table 1**). Among the protein-coding genes, *atpA*, *atpB*, and *atpE* play crucial roles in ATP synthesis, with GC content ranging from 36.6% to 40.5%, reflecting their conserved function in energy production. Photosystem-related genes such as *psaA*, *psaB*, *psbA*, and *psbB* exhibit GC content above 40%, indicating structural stability necessary for efficient electron transport. The ribosomal protein genes (*rpl*, *rps*) are distributed across the genome, with *rpl2*, *rpl14*, and *rpl23* being present in duplicate within the inverted repeat (IR) regions, supporting genome stability. Additionally, genes encoding NADH dehydrogenase subunits (*ndhA–ndhK*) are present, with *ndhF* being the longest at 2,247 bp, suggesting its significant role in chlororespiration. The genome also contains RNA polymerase genes (*rpoA*, *rpoB*, *rpoC1*, *rpoC2*), essential for plastid transcription, with *rpoC2* being the largest at 4,209 bp. The

presence of *matK*, a highly variable gene often used in phylogenetic studies, highlights the potential of this genome for evolutionary comparisons within Fabaceae.

The plastid genome of *C. ensiformis* also contains an extensive set of tRNA and rRNA genes, facilitating protein synthesis and plastid maintenance. A total of 41 tRNA genes, including *trnA-UGC*, *trnI-GAU*, and *trnK-UUU*, are distributed across the genome, with some occurring in duplicate within the IR regions. The rRNA genes (*rrn16*, *rrn23*, *rrn4.5*, and *rrn5*) are also duplicated in the IR regions, contributing to the structural conservation of the plastome. AT and GC skewness values indicate a strand asymmetry in nucleotide composition, with genes such as *psbL* and *ndhD* showing distinct biases. Start codon variation is observed, with most genes utilizing ATG, while *psbL* and *ndhD* use ACG, which may affect translational regulation. Additionally, genes such as *ycf1* and *ycf2*, which are among the longest in the genome, have an unclear function but are often associated with plastid genome stability. The overall organization of the *C. ensiformis* plastid genome, including its gene content and structural features, underscores its evolutionary adaptation and functional conservation within legumes.

The relative synonymous codon usage (RSCU) analysis of the *Canavalia ensiformis* plastid genome reveals significant codon preference patterns, reflecting both mutational biases and translational efficiency (**Fig. 2**). Codons ending in A or U are generally more frequent, consistent with the AT-rich nature of the plastid genome, which has a GC content of approximately 35%. Among amino acids, leucine, serine, and arginine exhibit the highest codon redundancy, as they are encoded by six different codons, with UUA and CUA being the most preferred for leucine. Similarly, codons for proline (*CCU*), threonine (*ACU*), and alanine (*GCU*) show a bias toward A/U-ending variants, suggesting selection for efficient translation in the plastid. Methionine (*AUG*) and tryptophan (*UGG*) are uniquely encoded by a single codon, maintaining strict conservation due to their essential functional roles. The presence of overrepresented codons such as *GAA* for glutamate and *UUU* for phenylalanine suggests adaptation to tRNA abundance and translational selection. These biases align with observations in other leguminous plastomes, where the prevalence of A/U-rich codons correlates with the availability of corresponding tRNAs.

Further analysis indicates that codon usage bias is influenced by both mutational pressures and functional constraints. Rarely used codons, particularly those ending in G or C, such as *CGG* for arginine and *GCG* for alanine, are underrepresented, likely due to the lower GC content of the genome. This bias is particularly evident in amino acids like glycine and valine, where *GGU* and *GUU* are favored over their GC-ending counterparts. The RSCU values also suggest translational optimization, with highly expressed genes preferentially using codons that correspond to abundant tRNAs, thereby improving translational efficiency. This trend is especially prominent in genes encoding essential photosynthetic proteins, such as *rbcL* and *psbA*, where efficient translation is critical for plastid function. Additionally, the bias toward A/U-rich codons may reduce the risk of secondary structure formation in mRNA, facilitating more efficient ribosome binding and elongation. Overall, the codon usage pattern of *C. ensiformis* reflects a balance between genome composition, translational selection, and evolutionary adaptation within the Fabaceae family.

The distribution of repeats in the *Canavalia ensiformis* plastid genome shows distinct patterns based on orientation, genomic regions, and length categories. Direct (D) repeats are consistently more frequent than palindrome (P) repeats across all regions, indicating a preference for direct repeat formations (**Fig. 3**). The **long single copy (LSC) region** contains the highest number of repeats, especially in the 25–29 and 30–39 length

categories, suggesting that this region is more prone to repeat formations. The **inverted repeat (IR) region**, which combines IRA and IRB, exhibits significantly fewer repeats than LSC but more than SSC, likely due to its duplicated nature stabilizing certain repeat structures. In the **short single copy (SSC) region**, repeats are the least frequent, with no direct repeats observed and only a few palindrome repeats present. This pattern suggests that the SSC region may be more conserved or less prone to repetitive sequences. Across all regions, the number of repeats declines as length increases, with the >40 category having the fewest occurrences.

When comparing orientation, direct (D) repeats are more evenly distributed across regions, while palindrome (P) repeats show more variation. The LSC region contains the highest number of both orientations, with **palindrome repeats more concentrated in shorter lengths**. The IR region maintains a relatively balanced distribution between direct and palindrome orientations, though it exhibits fewer overall repeats. In SSC, palindrome repeats are sporadic, reinforcing the idea that this region may be less involved in repeat expansion. The strong presence of short direct repeats in LSC and IR suggests these elements could play a role in genomic recombination or structural variation. The general decline of repeat frequency with increasing length indicates selective constraints against longer repeats. This pattern highlights the influence of genomic structure on repeat distribution and the potential functional roles of different repeat orientations in plastid genome stability.

The data presents the distribution of repeat motifs in the *Canavalia ensiformis* plastid genome, highlighting variations across the long single copy (LSC), inverted repeat (IR), and short single copy (SSC) regions (**Fig. 4**). The most frequent motifs appear to be A/T and AA/TT, particularly in the LSC and SSC regions, indicating a strong preference for simple sequence repeats composed of adenine and thymine. In contrast, motifs like C/G and CC/GG occur at significantly lower frequencies across all regions, suggesting a lower presence of cytosine-guanine repeats. The LSC region consistently contains the highest number of repeats, followed by SSC, while IR has the least, implying that repeat elements may be less stable or less frequently generated in the IR region. Additionally, as the length of the repeat motifs increases, their overall frequency appears to decrease, with shorter repeats (di- and tetra-nucleotides) being more common than longer ones. This trend is likely due to the natural bias of slippage replication in short simple sequence repeats, which are more prone to duplication. The lower occurrence of longer repeats may indicate selective constraints that limit their expansion within the plastid genome.

The IR region exhibits a unique pattern, showing notable differences in motif abundance compared to the LSC and SSC regions. While short repeats such as A/T and AA/TT are still present, their frequency in the IR region is significantly reduced, potentially due to its structural role in genome stabilization. Certain longer motifs, like AAAG/CTTT and AATAAT/ATTATT, are exclusively found or more common in the IR, suggesting that these repeats may contribute to structural integrity or recombination events. Interestingly, repeats such as ATATAT/ATATAT are entirely absent from SSC, highlighting regional specificity in repeat distribution. The scarcity of longer repeats in the SSC and IR regions may indicate higher selective pressure against their expansion, likely to prevent genomic instability. The dominance of A/T-rich repeats across all regions aligns with the general A-T richness observed in plastid genomes, which impacts their structural and functional properties. Overall, the dataset reveals that repeat motifs are not evenly distributed, and their presence is influenced by repeat length, genomic location, and structural constraints.

Comparative Analysis of Plastid Genome Among Fabaceae Family

The phylogenetic tree presented in this study depicts the evolutionary relationships among various species within the Fabaceae family. The comparative phylogeny, represented by Bayesian Inference and Maximum Likelihood, indicate a high degree of congruence between the two topologies, reinforcing the reliability of the inferred evolutionary relationships (**Fig. 5**). The phylogenetic analysis reveals the evolutionary placement of *Canavalia* within the Fabaceae family, with a well-supported monophyletic clade encompassing multiple *Canavalia* species. The tree topology indicates that *Canavalia* forms a distinct lineage closely related to other Fabaceae genera, such as *Hardwickia* and *Spatholobus*. The bootstrap values at the nodes defining the *Canavalia* clade are high, suggesting strong statistical support for its monophyly and evolutionary distinctiveness.

Within the *Canavalia* genus, species such as *Canavalia gladiata*, *Canavalia ensiformis*, *Canavalia rosea*, and *Canavalia cathartica* cluster together, reflecting their close genetic relationship. Notably, *Canavalia ensiformis* and *Canavalia gladiata*, both known for their economic and ecological significance, exhibit a closer evolutionary affinity, which aligns with previous studies suggesting their divergence from a common ancestor. The branching patterns further indicate that *Canavalia rosea*, a coastal and salt-tolerant species, may have diverged earlier from its terrestrial counterparts, possibly due to adaptive evolutionary pressures. The mirrored tree structure suggests that the phylogenetic relationships of *Canavalia* remain consistent across different analytical methods, reinforcing the reliability of these findings. However, minor topological variations between the two trees may indicate potential gene-tree discordance, which could result from hybridization events or incomplete lineage sorting. These discrepancies highlight the need for further studies incorporating additional molecular markers or whole-genome sequencing to refine the phylogenetic placement of *Canavalia* species.

The nucleotide variability (P_i) analysis of the *Canavalia ensiformis* plastid genome reveals substantial variation across coding sequences (CDS), tRNAs, and rRNAs, indicating regions of evolutionary conservation and divergence (**Fig. 6**). Among CDS genes, *ycf1* and *ycf1b* exhibit the highest P_i values (0.128), suggesting their roles as highly variable regions, which are often associated with adaptive evolution. The *accD* gene also shows relatively high variability (0.118), consistent with its involvement in plastid gene regulation and possible selective pressures. Conversely, genes like *psbL* (0.0136) and *psbF* (0.0228) exhibit lower variability, indicating evolutionary conservation likely due to their essential roles in photosynthetic processes. The ATP synthase subunit genes (*atpA*, *atpB*, *atpE*, *atpF*, *atpH*, *atpI*) show moderate P_i values ranging from 0.036 to 0.051, reflecting functional constraints balanced with some degree of mutational flexibility. Similarly, ribosomal proteins such as *rpl32* (0.0781) and *rps15* (0.0783) display higher nucleotide diversity, suggesting their potential for structural or functional modifications. Overall, CDS regions with higher P_i values may serve as markers for phylogenetic studies and evolutionary adaptation analysis within Fabaceae.

In contrast, the tRNA and rRNA genes show significantly lower nucleotide variability, highlighting their structural and functional conservation. Many tRNA genes, including *trnI-GAU*, *trnI-UAG*, and *trnR-ACG*, exhibit zero P_i values, indicating complete conservation due to their crucial role in translation fidelity. Among tRNAs with moderate variability, *trnM-CAU* (0.0225) and *trnQ-UUG* (0.0174) show relatively higher diversity, suggesting potential sequence modifications for optimal codon recognition. The ribosomal RNA genes (*rnm16*,

rrn23, *rrn4.5*, *rrn5*) also exhibit low Pi values (ranging from 0.0045 to 0.0158), reinforcing their functional constraint in maintaining ribosomal integrity. The relatively higher variability of *rrn4.5* (0.0158) may reflect minor structural adaptations within the plastid ribosome. These findings indicate that while CDS regions exhibit a range of nucleotide diversity due to functional divergence and evolutionary pressures, tRNA and rRNA genes remain highly conserved to preserve translational and ribosomal efficiency. Consequently, nucleotide variability data can be valuable for identifying regions under selective pressure and understanding the evolutionary dynamics of the *C. ensiformis* plastome.

Discussion

This study provides the first comprehensive analysis of the *Canavalia ensiformis* plastid genome, revealing a total length of 157,809 bp with a quadripartite structure typical of plastid genomes in angiosperms. The genome comprises 136 genes, including protein-coding genes involved in essential functions such as photosynthesis (e.g., *psaA*, *psbA*, *rbcL*), ATP synthesis (*atpA*, *atpB*), and transcription (*rpoA*, *rpoB*). Notably, the presence of *ndh* genes suggests an alternative electron transport mechanism, which may support photosynthesis under stress conditions. Codon usage analysis highlighted a preference for A/U-rich codons, reflecting translational efficiency and mutational biases in the genome. Additionally, repeat sequence analysis revealed a high frequency of direct repeats in the LSC region, contributing to potential recombination events and genome evolution. Comparative analyses with other Fabaceae species positioned *C. ensiformis* closely with *Canavalia gladiata*, reinforcing its evolutionary relationships within the legume family.

The structural organization and gene content of the *C. ensiformis* plastid genome align with previous reports on leguminous plastomes. Similar genome sizes and quadripartite structures have been observed in other Fabaceae members, such as *Glycine max* (soybean) and *Phaseolus vulgaris* (common bean), with gene numbers remaining largely conserved. However, the expansion of the IR region in *C. ensiformis* is notable, as variations in IR length can influence genome stability and recombination rates. The codon usage bias toward A/U-rich codons is consistent with findings in other legume plastomes, supporting the hypothesis that these preferences enhance translational efficiency. Moreover, phylogenetic analyses confirmed *Canavalia* species as a distinct lineage within Fabaceae, aligning with molecular markers used in previous studies. The presence of highly conserved photosynthetic and ribosomal genes reinforces the functional stability of the plastid genome, a pattern also documented in related legumes.

The plastid genome of *C. ensiformis* encodes a complete set of genes required for photosynthesis, transcription, and translation, emphasizing its role in sustaining autotrophic growth. The presence of the *ndh* gene family suggests a mechanism for optimizing electron transport under fluctuating environmental conditions, which could enhance the plant's adaptability to diverse habitats. Additionally, the retention of essential genes such as *rbcL* and *psbA* ensures robust carbon fixation and efficient light energy utilization. The observed codon bias toward A/U-rich codons may indicate a selection pressure for efficient protein synthesis, potentially reducing metabolic costs in translation. Furthermore, repeat elements within the LSC region could contribute to genome plasticity, affecting recombination dynamics and evolutionary adaptation. These features collectively highlight the evolutionary conservation and functional resilience of the *C. ensiformis* plastid genome within Fabaceae.

Despite providing valuable insights, this study has certain limitations. First, while comparative genomic analyses identified key similarities and differences among Fabaceae plastomes, a broader taxonomic sampling could provide a more comprehensive understanding of phylogenetic relationships. Second, functional validation of hypothetical genes (*ycf1*, *ycf2*) remains challenging, as their roles in plastid biogenesis and stability are still unclear. Additionally, while codon usage bias was analyzed, its direct impact on gene expression and protein synthesis was not experimentally validated. Lastly, environmental factors influencing plastid genome evolution, such as selective pressures from habitat adaptation, were not explored in detail. Future studies incorporating transcriptomic and proteomic approaches could help address these gaps and provide a more functional perspective on plastid genome evolution in Fabaceae.

Future research should focus on expanding comparative genomic analyses across a wider range of Fabaceae species to better understand evolutionary patterns and structural variations in plastid genomes. Functional studies on hypothetical genes such as *ycf1* and *ycf2* could elucidate their roles in plastid development and stability. Additionally, experimental validation of codon bias effects on gene expression through transcriptomic studies could provide deeper insights into translational efficiency and metabolic regulation. Investigating the impact of environmental stressors on plastid genome stability, particularly in relation to repeat sequence expansion, could further clarify adaptive mechanisms in legumes. Finally, incorporating plastid genome editing techniques could help explore gene function and enhance crop improvement strategies for Fabaceae species, particularly for stress tolerance and photosynthetic efficiency.

This study presents a comprehensive characterization of the *C. ensiformis* plastid genome, reinforcing its structural conservation and functional significance within Fabaceae. The plastid genome exhibits features that contribute to photosynthetic efficiency, genome stability, and evolutionary adaptation. Comparative analyses confirm its close phylogenetic relationship with other *Canavalia* species and highlight conserved plastid genome features within legumes. Despite some limitations, this study provides a foundational reference for future genomic and evolutionary research on *C. ensiformis* and related Fabaceae species. By advancing our understanding of plastid genome evolution and function, this research contributes valuable insights for plant genetics, phylogenomics, and potential agricultural applications.

Declarations

Acknowledgements

The research was funded by the 'Research and Innovation for Advanced Indonesia (Riset dan Inovasi untuk Indonesia Maju, RIIM)' grant batch IV from the Indonesia Endowment Fund for Education (LPDP) and the National Research and Innovation Agency (BRIN), according to contract number 152/IV/KS/11/2023 and PKS/206/XI/2023/UNHANRI.

Author contributions statement

I.A.N and R.D.S conceived the experiments; I.A.N performed DNA genome isolation and quality assessment; R.D.S. conducted the data curation, bioinformatics, and statistical analysis; I.A.N. and R.D.S. wrote the original draft preparation; W.A.M., W.H.A., and R. revised the manuscript; All authors reviewed the manuscript.

Data Availability

Data are included in the manuscript or supplementary information files. The raw reads were submitted to NCBI under ID: PRJNA1227101. The assembled chloroplast is accessible via accession number PV174432.

Competing interests

The authors declare no competing interests.

References

1. Akande, K. E. (2016). *Effects of anti-nutritive constituents of unprocessed jack bean (Canavalia ensiformis)(L)(DC) seed meal on poultry performance.*
2. Anitha, K., Mohana Lakshmi, S., & Satyanarayana, S. V. (2020). Antidiabetic, lipid lowering and antioxidant potentiating effect of Canavalia species in high fat diet-streptozotocin induced model. *Advances in Traditional Medicine*, 20, 609–618.
3. Balliana, A. G., Moura, B. B., Inckot, R. C., & Bona, C. (2017). Development of Canavalia ensiformis in soil contaminated with diesel oil. *Environmental Science and Pollution Research*, 24(1), 979–986. <https://doi.org/10.1007/s11356-016-7674-1>
4. Barney, J. N., & DiTomaso, J. M. (2010). Bioclimatic predictions of habitat suitability for the biofuel switchgrass in North America under current and future climate scenarios. *Biomass and Bioenergy*, 34(1), 124–133. <https://doi.org/https://doi.org/10.1016/j.biombioe.2009.10.009>
5. Bhat, J. A., Ali, S., Salgotra, R. K., Mir, Z. A., Dutta, S., Jadon, V., Tyagi, A., Mushtaq, M., Jain, N., & Singh, P. K. (2016). Genomic Selection in the Era of Next Generation Sequencing for Complex Traits in Plant Breeding. *Frontiers in Genetics*, 7, 1–11. <https://doi.org/10.3389/fgene.2016.00221>
6. Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, 23(19), 2633–2635. <https://doi.org/10.1093/bioinformatics/btm308>
7. Breseghello, F., & Coelho, A. S. G. (2013). Traditional and Modern Plant Breeding Methods with Examples in Rice (*Oryza sativa* L.). *Journal of Agricultural and Food Chemistry*, 61(35), 8277–8286. <https://doi.org/10.1021/jf305531j>
8. Carlini, C. R., & Guimaraes, J. A. (2001). Canatoxin, a highly toxic protein from Canavalia ensiformis. *Arquivos de Biologia e Tecnologia*, 23(2).
9. Chaudhari, N., & Pathak, B. (2022). Assessment of Floristic Diversity and its Structural Composition in South Gujarat. *Indian Journal of Ecology*, 49(1), 64–74.
10. Chel-Guerrero, L., Pérez-Flores, V., Betancur-Ancona, D., & Dávila-Ortiz, G. (2002). Functional Properties of Flours and Protein Isolates from Phaseolus lunatus and Canavalia ensiformis Seeds. *Journal of Agricultural and Food Chemistry*, 50(3), 584–591. <https://doi.org/10.1021/jf010778j>
11. Clement, K., Rees, H., Canver, M. C., Gehrke, J. M., Farouni, R., Hsu, J. Y., Cole, M. A., Liu, D. R., Joung, J. K., Bauer, D. E., & Pinello, L. (2019). CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nature Biotechnology*, 37(3), 224–226. <https://doi.org/10.1038/s41587-019-0032-3>
12. D’Cunha, M., & Sridhar, K. R. (2010). L-canavanine and L-arginine in two wild legumes of the genus Canavalia. *Inst Integr Omics Appl Biotechnol J*, 1, 29–33.

13. de Baan, L., Curran, M., Rondinini, C., Visconti, P., Hellweg, S., & Koellner, T. (2015). High-Resolution Assessment of Land Use Impacts on Biodiversity in Life Cycle Assessment Using Species Habitat Suitability Models. *Environmental Science & Technology*, 49(4), 2237–2244. <https://doi.org/10.1021/es504380t>
14. Dhawan, S. S., Rai, G. K., Darokar, M. P., Lal, R. K., Misra, H. O., & Khanuja, S. P. S. (2011). Comparative genetic analysis of trichome-less and normal pod genotypes of *Mucuna pruriens* (Fabaceae). *Genetics and Molecular Research*, 10(3), 2049–2056. <https://doi.org/10.4238/vol10-3gmr1264>
15. Díaz, L. A. G. (2014). *Ensilage of Jack bean and cowpea grains sole or mixed with sorghum to improve their nutritional value as feedstuff for growing-finishing pigs*. Universität Rostock.
16. Dierckxsens, N., Mardulyn, P., & Smits, G. (2017). NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research*, 45(4), e18–e18.
17. Doss, A., Pugalenth, M., Vadivel, V. G., Subhashini, G., & Anitha Subash, R. (2011). Effects of processing technique on the nutritional composition and antinutrients content of under-utilized food legume *Canavalia ensiformis* L.DC. *International Food Research Journal*, 18(3), 965–970.
18. Duitama, J., Silva, A., Sanabria, Y., Cruz, D. F., Quintero, C., Ballen, C., Lorieux, M., Scheffler, B., Farmer, A., Torres, E., Oard, J., & Tohme, J. (2015). Whole Genome Sequencing of Elite Rice Cultivars as a Comprehensive Information Resource for Marker Assisted Selection. *PLOS ONE*, 10(4), e0124617. <https://doi.org/10.1371/journal.pone.0124617>
19. El-Kassaby, Y. A., Cappa, E. P., Liewlaksaneeyanawin, C., Klápště, J., & Lstibůrek, M. (2011). Breeding without Breeding: Is a Complete Pedigree Necessary for Efficient Breeding? *PLOS ONE*, 6(10), e25737. <https://doi.org/10.1371/journal.pone.0025737>
20. Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, 6(5), 1–10. <https://doi.org/10.1371/journal.pone.0019379>
21. Evans, J., McCormick, R. F., Morishige, D., Olson, S. N., Weers, B., Hilley, J., Klein, P., Rooney, W., & Mullet, J. (2013). Extensive variation in the density and distribution of DNA polymorphism in sorghum genomes. *PloS One*, 8(11), e79192. <https://doi.org/10.1371/journal.pone.0079192>
22. Faralli, M., & Lawson, T. (2020). Natural genetic variation in photosynthesis: an untapped resource to increase crop yield potential? *The Plant Journal*, 101(3), 518–528. <https://doi.org/https://doi.org/10.1111/tpj.14568>
23. Fernie, A. R., Tadmor, Y., & Zamir, D. (2006). Natural genetic variation for improving crop quality. *Current Opinion in Plant Biology*, 9(2), 196–202. <https://doi.org/https://doi.org/10.1016/j.pbi.2006.01.010>
24. Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37(12), 4302–4315.
25. Follmer, C., Barcellos, G. B. S., Zingali, R. B., Machado, O. L. T., Alves, E. W., Barja-Fidalgo, C., Guimarães, J. A., & Carlini, C. R. (2001). Canatoxin, a toxic protein from jack beans (*Canavalia ensiformis*), is a variant form of urease (EC 3.5.1.5): biological effects of urease independent of its ureolytic activity. *Biochemical Journal*, 360(1), 217–224. <https://doi.org/10.1042/bj3600217>
26. Fukasawa, Y., Ermini, L., Wang, H., Carty, K., & Cheung, M.-S. (2020). LongQC: a quality control tool for third generation sequencing long read data. *G3: Genes, Genomes, Genetics*, 10(4), 1193–1196.

27. Harvian, Z. A., Ningrum, A., Anggrahini, S. R. I., & Setyaningsih, W. (2019). In silico approach in evaluation of jack bean (*Canavalia ensiformis*) canavalin protein as precursors of bioactive peptides with dual antioxidant and angiotensin i-converting enzyme inhibitor. *Materials Science Forum*, 948, 85–94.
28. He, J., Zhao, X., Laroche, A., Lu, Z. X., Liu, H. K., & Li, Z. (2014). Genotyping-by-sequencing (GBS), An ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Frontiers in Plant Science*, 5(SEP), 1–8. <https://doi.org/10.3389/fpls.2014.00484>
29. Henry, R. J., & Nevo, E. (2014). Exploring natural selection to guide breeding for agriculture. *Plant Biotechnology Journal*, 12(6), 655–662. <https://doi.org/https://doi.org/10.1111/pbi.12215>
30. Henschel, R., Lieber, M., Wu, L.-S., Nista, P. M., Haas, B. J., & LeDuc, R. D. (2012). Trinity RNA-Seq assembler performance optimization. *Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the EXtreme to the Campus and Beyond*, 1–8.
31. Hijmans, R. J., Guarino, L., Cruz, M., & Rojas, E. (2001). Computer tools for spatial analysis of plant genetic resources data: 1. DIVA-GIS. *Plant Genetic Resources Newsletter*, 15–19.
32. Holm LG; Pancho JV; Herbenger JP; Plucknett DL, 1979. A Geographical Atlas of World Weeds. New York, USA: John Wiley & Sons.
33. ILDIS, 2014. International Legume Database and Information Service. Reading, UK: School of Plant Sciences, University of Reading. <http://www.ildis.org/>
34. Jiang, G., Zeng, J., & He, Y. (2014). Analysis of quantitative trait loci affecting chlorophyll content of rice leaves in a double haploid population and two backcross populations. *Gene*, 536(2), 287–295. <https://doi.org/10.1016/j.gene.2013.12.010>
35. Langmead, B. and, & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), 357–359. <https://doi.org/10.1038/nmeth.1923>
36. Leger, A., & Leonardi, T. (2019). pycoQC, interactive quality control for Oxford Nanopore Sequencing. *Journal of Open Source Software*, 4(34), 1236.
37. Lenkala, P., Rani, K. R., Sivaraj, N., Reddy, K. R., & Prada, M. J. (2015). Genetic variability and character association studies in Jack bean (*Canavalia ensiformis* (L.) DC.) for yield and yield contributing traits. *Electronic Journal of Plant Breeding*, 6(2), 625–629.
38. Letunic, I., & Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research*, 47(W1), 256–259. <https://doi.org/10.1093/nar/gkz239>
39. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
40. Liu, M., Ding, M., Chen, L., Ouyang, K., Hui, W., Li, J., & Chen, X. (2014). Genetic diversity and relationships among *Canavalia ensiformis* (L.) DC. Accessions as revealed by sequence-related amplified polymorphism markers. *Biochemical Systematics and Ecology*, 57, 242–249. <https://doi.org/https://doi.org/10.1016/j.bse.2014.08.005>
41. Liu, Y., Zhang, L. Y., & Li, J. (2019). Fast detection of maximal exact matches via fixed sampling of query K-mers and Bloom filtering of index K-mers. *Bioinformatics*, 35(22), 4560–4567.
42. Luitel, D. R., Siwakoti, M., Joshi, M. D., Rangaswami, M., & Jha, P. K. (2020). Potential suitable habitat of *Eleusine coracana* (L) gaertn (Finger millet) under the climate change scenarios in Nepal. *BMC Ecology*,

- 20(1), 19. <https://doi.org/10.1186/s12898-020-00287-6>
43. Lynch, M., & O'Hely, M. (2001). Captive breeding and the genetic fitness of natural populations. *Conservation Genetics*, 2(4), 363–378. <https://doi.org/10.1023/A:1012550620717>
 44. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution*, 38(10), 4647–4654.
 45. Martínez-Macias, K. J., Márquez-Guerrero, S. Y., Martínez-Sifuentes, A. R., & Segura-Castruita, M. Á. (2022). Habitat Suitability of Fig (*Ficus carica* L.) in Mexico under Current and Future Climates. In *Agriculture* (Vol. 12, Issue 11). <https://doi.org/10.3390/agriculture12111816>
 46. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20, 1297–1303. <https://doi.org/10.1101/gr.107524.110.20>
 47. Miernyk, J. A., & Hajdich, M. (2011). Seed proteomics. *Journal of Proteomics*, 74(4), 389–400.
 48. Morris, J. B. (2007). Swordbean (*Canavalia ensiformis* (L.) DC.) Genetic Resources Regenerated for Potential Medical, Nutraceutical and Agricultural Traits. *Genetic Resources and Crop Evolution*, 54(3), 585–592. <https://doi.org/10.1007/s10722-006-0015-3>
 49. Patel, R. K., & Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PloS One*, 7(2), e30619.
 50. Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E., & Blair, M. E. (2017). Opening the black box: An open-source release of Maxent. *Ecography*, 40(7), 887–893.
 51. Phillips, S. J., & Dudík, M. (2008). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, 31(2), 161–175.
 52. Pier, 2014. Pacific Islands Ecosystems at Risk. Honolulu, USA: HEAR, University of Hawaii. <http://www.hear.org/pier/index.html>
 53. Postal, M., Martinelli, A. H. S., Becker-Ritt, A. B., Ligabue-Braun, R., Demartini, D. R., Ribeiro, S. F. F., Pasquali, G., Gomes, V. M., & Carlini, C. R. (2012). Antifungal properties of *Canavalia ensiformis* urease and derived peptides. *Peptides*, 38(1), 22–32.
 54. Postma, E. (2006). Implications of the difference between true and predicted breeding values for the study of natural selection and micro-evolution. *Journal of Evolutionary Biology*, 19(2), 309–320. <https://doi.org/https://doi.org/10.1111/j.1420-9101.2005.01007.x>
 55. Pugalenth, M., Doss, A., Rajendrakumaran, D., & Vadivel, V. (2010). Antibacterial activity of *Canavalia ensiformis* L. DC seeds due to solvent extract and fraction method. *Indian Journal of Science and Technology*, 3(1), 94–95.
 56. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3), 559–575. <https://doi.org/10.1086/519795>
 57. Qi, A., Keatinge, J. D. H., Wheeler, T. R., Papastylianou, I., Subedi, M., Shah, P. B., Musitwa, F., Cespedes, E., Bening, C., Ellis, R. H., & Summerfield, R. J. (2000). Validation of a Photothermal Phenology Model for

- Predicting Dates of Flowering and Maturity in Legume Cover Crops using Field Observations. *Biological Agriculture & Horticulture*, 17(4), 349–365. <https://doi.org/10.1080/01448765.2000.9754855>
58. Ramli, N. A. M., Chen, Y. H., Zin, Z. M., Abdullah, M. A. A., Rusli, N. D., & Zainol, M. K. (2021). Effect of soaking time and fermentation on the nutrient and antinutrients composition of *Canavalia ensiformis* (Kacang Koro). *IOP Conference Series: Earth and Environmental Science*, 756(1), 12033.
 59. Randall RP, 2012. A Global Compendium of Weeds. Perth, Australia: Department of Agriculture and Food Western Australia, 1124 pp.
 60. Rifai MA. 2015. *Asas-Asas Sistematika Biologi*. Bogor (ID): Herbarium Bogoriense Puslit Biologi LIPI.
 61. Robertson, T., Döring, M., Guralnick, R., Bloom, D., Wieczorek, J., Braak, K., Otegui, J., Russell, L., & Desmet, P. (2014). The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet. *PLOS ONE*, 9(8), e102623. <https://doi.org/10.1371/journal.pone.0102623>
 62. Rugayah, Retnowati A, Windadri FI, Hidayat A. 2004. Pengumpulan data taksonomi. In: Rugayah, Widjaja EA, Praptiwi, editor. *Pedoman Pengumpulan Data Keanekaragaman Flora*. Bogor (ID): Pusat Penelitian Biologi-Lembaga Ilmu Pengetahuan Indonesia.
 63. Saghai-Maroo, M. A., Soliman, K. M., Jorgensen, R. A., & Allard, R. W. (1984). Ribosomal DNA spacer-length polymorphisms in barley: mendelian inheritance, chromosomal location, and population dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, 81(24), 8014–8018. <https://doi.org/10.1073/pnas.81.24.8014>
 64. Savadi, S., Mangalassery, S., & Sandesh, M. S. (2021). Advances in genomics and genome editing for breeding next generation of fruit and nut crops. *Genomics*, 113(6), 3718–3734. <https://doi.org/https://doi.org/10.1016/j.ygeno.2021.09.001>
 65. Shahinnia, F., Tricker, P. J., Hajirezaei, M., & Chen, Z. (2019). *Genomics Assisted Breeding of Crops for Abiotic Stress Tolerance, Vol. II* (Vol. 21). <https://doi.org/10.1007/978-3-319-99573-1>
 66. Sinn, B. T., Simon, S. J., Santee, M. V, DiFazio, S. P., Fama, N. M., & Barrett, C. F. (2022). ISSRseq: An extensible method for reduced representation sequencing. *Methods in Ecology and Evolution*, 13(3), 668–681. <https://doi.org/https://doi.org/10.1111/2041-210X.13784>
 67. Staszek, P., Weston, L. A., Ciacka, K., Krasuska, U., & Gniazdowska, A. (2017). L-Canavanine: How does a simple non-protein amino acid inhibit cellular function in a diverse living system? *Phytochemistry Reviews*, 16, 1269–1282.
 68. Suvarna, G., Prabhu, A., & Muhseena, K. (2020). Cytotoxic and Apoptotic Efficacy of Concanavalins Isolated from *Canavalia ensiformis* and *Canavalia gladiata* on Colon Carcinoma Cells. *Indian Journal of Pharmaceutical Sciences*, 82(6), 967–971.
 69. Telenius, A. (2011). Biodiversity information goes public: GBIF at your service. *Nordic Journal of Botany*, 29(3), 378–381. <https://doi.org/https://doi.org/10.1111/j.1756-1051.2011.01167.x>
 70. Teodoro, M. S., Castro, K. N. D. E. C., & Magalhães, J. A. (2018). Assessment of legumes with potential use as green manure in the Coastal Tablelands of Piauí State, Brazil. In *Revista Caatinga* (Vol. 31). scielo.
 71. Trijatmiko, K. R., Supriyanta, Prasetyono, J., Thomson, M. J., Vera Cruz, C. M., Moeljopawiro, S., & Pereira, A. (2014). Meta-analysis of quantitative trait loci for grain yield and component traits under reproductive-stage drought stress in an upland rice population. *Molecular Breeding*, 34(2), 283–295. <https://doi.org/10.1007/s11032-013-0012-0>

72. USDA-ARS, 2014. Germplasm Resources Information Network (GRIN). Online Database. Beltsville, Maryland, USA: National Germplasm Resources Laboratory. <https://npgsweb.ars-grin.gov/gringlobal/taxon/taxonomysearch.aspx>

73. Waterhouse BM; Mitchell AA, 1998. Northern Australia Quarantine Strategy: weeds target list. Australian Quarantine & Inspection Service, 110 pp. [Miscellaneous Publication No. 6/98.]

74. Zimin, A. V, Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., & Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics*, 29(21), 2669–2677.

Table

Table 1. Gene list for the annotated jack bean (*Canavalia ensiformis*) plastid genome

Feature*	Strand	Size (bp)	Start	End	GC(%)	AT skewness	GC skewness	Start Codon	Terminal Codon
accD	+	1575	27223	28797	32,1	-0,002	0,192	ATG	TAA
atpA	-	1533	6759	8291	39,1	0,066	0,085	ATG	TGA
atpB	+	1476	135933	137408	40,5	0,042	0,072	ATG	TAG
atpE	+	402	137431	137832	39,5	0,136	0,145	ATG	TAA
atpF	-	636	8406	9752	36,6	0,102	0,245	ATG	TAG
atpH	-	246	10244	10489	44,3	-0,168	0,193	ATG	TAA
atpI	-	744	11165	11908	34,4	-0,156	0,031	ATG	TGA
ccsA	+	942	81495	82436	29,3	-0,084	0,109	ATG	TGA
cemA	+	690	30604	31293	30,2	-0,083	-0,038	ATG	TGA
clpP1	-	597	39004	41103	40,8	-0,040	0,177	ATG	TAA
matK	-	1524	130477	132000	29,0	-0,111	0,023	ATG	TAA
ndhA	-	1104	86995	89350	32,3	-0,143	0,053	ATG	TAA
ndhB	-	1479	63352	65515	37,5	-0,121	-0,052	ATG	TAG
ndhB_2	+	1479	108928	111091	37,5	-0,121	-0,052	ATG	TAG
ndhC	+	363	139250	139612	33,9	-0,258	0,187	ATG	TAA
ndhD	-	1497	82864	84360	33,9	-0,160	0,026	ACG	TAG
ndhE	-	306	85029	85334	32,7	-0,126	0,020	ATG	TAG
ndhF	-	2247	77702	79948	29,8	-0,173	0,081	ATG	TAG
ndhG	-	531	85603	86133	32,4	-0,198	0,035	ATG	TAA
ndhH	-	1182	89352	90533	37,4	0,019	0,213	ATG	TGA
ndhI	-	501	86421	86921	32,2	-0,012	0,068	ATG	TAA
ndhJ	+	477	140469	140945	37,2	-0,013	0,119	ATG	TAA
ndhK	+	669	139664	140332	38,7	0,020	-0,050	ATG	TGA
petA	+	963	31522	32484	38,3	0,007	0,111	ATG	TAG
petB	+	648	44120	45571	38,4	-0,218	0,124	ATG	TAG
petD	+	483	45767	46970	37,7	-0,090	0,044	ATG	TAA
petG	+	114	34909	35022	35,9	-0,288	0,220	ATG	TGA
petL	+	96	34639	34734	32,3	-0,231	-0,032	ATG	TGA
petN	+	90	25557	25646	35,5	-0,276	0,312	ATG	TAA

psaA	+	2253	147993	150245	41,1	-0,101	0,036	ATG	TAA
psaB	+	2205	150271	152475	40,1	-0,145	0,063	ATG	TAA
psaC	-	246	84483	84728	42,3	0,000	0,154	ATG	TGA
psaI	+	105	29263	29367	32,3	-0,211	-0,176	ATG	TAG
psaJ	+	135	35943	36077	39,3	-0,268	0,019	ATG	TAG
psbA	-	1062	128762	129823	40,5	-0,201	0,081	ATG	TAA
psbB	+	1527	41617	43143	42,9	-0,163	0,197	ATG	TAA
psbC	-	1422	154849	156270	41,9	-0,167	0,141	ATG	TGA
psbD	-	1062	156218	157279	42,2	-0,235	0,062	ATG	TAA
psbE	-	252	33430	33681	38,5	-0,123	0,052	ATG	TAG
psbF	-	120	33301	33420	43,3	-0,088	-0,038	ATG	TAA
psbH	+	222	43766	43987	36,5	-0,106	0,210	ATG	TAG
psbI	+	111	4634	4744	36,0	-0,183	-0,100	ATG	TAA
psbJ	-	123	32876	32998	39,9	-0,297	0,143	ATG	TAG
psbK	+	186	3970	4155	34,4	-0,361	-0,062	ATG	TGA
psbL	-	117	33162	33278	31,6	-0,075	-0,189	ACG	TAA
psbM	+	105	2311	2415	32,3	-0,099	-0,059	ATG	TAA
psbN	-	132	43518	43649	43,2	-0,040	-0,018	ATG	TAG
psbT	+	108	43353	43460	32,4	-0,096	-0,086	ATG	TGA
psbZ	-	189	153948	154136	31,2	-0,338	0,085	ATG	TGA
rbcL	-	1428	133760	135187	42,3	-0,039	0,142	ATG	TAA
rpl14	-	369	48893	49261	37,2	0,103	0,124	ATG	TAA
rpl14_2	+	369	125182	125550	37,2	0,103	0,124	ATG	TAA
rpl16	-	408	49388	50954	43,9	0,188	0,084	ATG	TAA
rpl2	-	822	53021	54525	42,5	0,175	0,152	ATG	TAG
rpl2_2	+	822	119918	121422	42,5	0,175	0,152	ATG	TAG
rpl20	-	393	37482	37874	31,5	0,167	0,081	ATG	TAA
rpl23	-	282	54544	54825	37,2	0,153	0,124	ATG	TAA
rpl23_2	+	282	119618	119899	37,2	0,153	0,124	ATG	TAG
rpl32	+	162	80686	80847	30,8	0,161	0,040	ATG	TAA
rpl33	+	201	36571	36771	34,3	0,152	0,188	ATG	TAG

rpl36	-	114	47928	48041	37,7	0,239	0,163	ATG	TAA
rpl36_2	+	114	126402	126515	37,7	0,239	0,163	ATG	TAA
rpoA	+	1002	127410	128411	32,3	0,108	0,015	ATG	TAA
rpoB	-	3261	20290	23550	37,1	0,056	0,120	ATG	TAA
rpoC1	-	2055	17433	20259	36,3	0,037	0,130	ATG	TAA
rpoC2	-	4209	13043	17251	34,4	0,068	0,070	ATG	TAA
rps11	-	450	47108	47557	40,9	0,195	0,185	ATG	TAA
rps11_2	+	450	126886	127335	40,9	0,195	0,185	ATG	TAA
rps12	-	384	38624	67204	43,0	0,178	0,006	ATG	TGA
rps14	+	303	152617	152919	39,6	0,137	0,217	ATG	TAA
rps15	-	330	90636	90965	28,5	0,127	0,085	ATG	TAA
rps16	+	270	26003	27100	34,5	0,085	0,118	ATG	TAA
rps18	+	429	36968	37396	31,0	0,176	-0,008	ATG	TAA
rps19	-	279	52293	52571	32,2	0,217	-0,111	ATG	TAA
rps19_2	+	279	121872	122150	32,2	0,217	-0,111	ATG	TAA
rps2	-	711	12145	12855	36,4	0,080	0,120	ATG	TGA
rps3	-	657	51109	51765	32,3	0,200	0,123	ATG	TAA
rps3_2	+	657	122678	123334	32,3	0,200	0,123	ATG	TAA
rps4	+	597	144156	144752	35,9	0,097	-0,047	ATG	TAA
rps7	-	468	65876	66343	40,2	0,214	0,053	ATG	TAA
rps7_2	+	468	108100	108567	40,2	0,214	0,053	ATG	TAA
rps8	-	405	48224	48628	33,8	0,179	0,153	ATG	TAA
rps8_2	+	405	125815	126219	33,8	0,179	0,153	ATG	TAA
ycf1	-	5268	91312	96579	27,2	0,137	0,054	ATG	TAA
ycf2	-	6717	112340	119056	36,6	-0,007	0,022	ATG	TAA
ycf2_2	+	6717	55387	62103	36,6	-0,007	0,022	ATG	TAA
ycf3	+	507	145320	147263	38,2	0,067	0,041	ATG	TAA
ycf4	+	585	29576	30160	34,2	-0,143	0,200	ATG	TGA
rrn16	+	1491	69313	70803	55,9	0,139	0,151	-	-
rrn16_2	-	1491	103640	105130	55,9	0,139	0,151	-	-
rrn23	+	2813	73223	76035	54,9	0,178	0,147	-	-

rrn23_2	-	2813	98408	101220	54,9	0,178	0,147	-	-
rrn4.5	+	104	76134	76237	51,0	0,137	0,208	-	-
rrn4.5_2	-	104	98206	98309	51,0	0,137	0,208	-	-
rrn5	+	121	76459	76579	52,9	0,053	0,094	-	-
rrn5_2	-	121	97864	97984	52,9	0,053	0,094	-	-
trnA-UGC	+	892	72171	73062	49,4	0,011	0,129	-	-
trnA-UGC_2	-	892	101381	102272	49,4	0,011	0,129	-	-
trnC-GCA	+	75	24656	24730	58,6	-0,161	0,136	-	-
trnD-GUC	+	74	1810	1883	62,2	-0,071	0,087	-	-
trnE-UUC	+	73	1137	1209	58,9	-0,200	0,023	-	-
trnF-GAA	-	75	141272	141346	53,3	0,029	0,100	-	-
trnG-GCC	-	71	153294	153364	50,7	-0,029	0,056	-	-
trnG-UCC	-	71	153294	153364	50,7	-0,029	0,056	-	-
trnH-GUG	-	75	128414	128488	56,0	-0,091	0,143	-	-
trnI-CAU	-	74	55224	55297	46,0	0,000	0,000	-	-
trnI-CAU_2	+	74	119146	119219	46,0	0,000	0,000	-	-
trnI-GAU	+	1009	71098	72106	48,3	-0,038	0,047	-	-
trnI-GAU_2	-	1009	102337	103345	48,3	-0,038	0,047	-	-
trnI-UAG	+	1010	71097	72106	48,3	-0,038	0,045	-	-
trnI-UAG_2	-	1010	102337	103346	48,3	-0,038	0,045	-	-
trnK-UUU	-	2652	130095	132746	29,8	-0,087	0,078	-	-
trnL-CAA	-	81	62687	62767	51,9	0,077	0,190	-	-
trnL-CAA_2	+	81	111676	111756	51,9	0,077	0,190	-	-

trnL-UAA	-	610	141667	142276	34,0	0,141	0,053	-	-
trnL-UAG	+	80	81310	81389	56,3	-0,029	0,200	-	-
trnM-CAU	-	73	137938	138010	41,1	-0,070	0,133	-	-
trnN-GUU	-	72	77550	77621	54,1	-0,212	0,231	-	-
trnN-GUU_2	+	72	96822	96893	54,1	-0,212	0,231	-	-
trnP-GGG	-	71	35433	35503	50,7	-0,200	0,167	-	-
trnP-UGG	-	74	35431	35504	48,7	-0,158	0,167	-	-
trnQ-UUG	-	72	3390	3461	62,5	-0,185	0,111	-	-
trnR-ACG	+	74	76854	76927	62,1	0,000	0,043	-	-
trnR-ACG_2	-	74	97516	97589	62,1	0,000	0,043	-	-
trnR-UCU	+	72	6227	6298	43,0	-0,024	0,097	-	-
trnS-GAA	-	90	144982	145071	50,0	-0,244	0,200	-	-
trnS-GCU	-	87	4770	4856	49,4	-0,182	0,163	-	-
trnS-GGA	-	88	144982	145069	51,2	-0,256	0,200	-	-
trnS-UGA	+	93	154528	154620	50,6	0,000	0,106	-	-
trnT-GGU	-	72	655	726	48,6	0,027	0,143	-	-
trnT-UGU	+	73	143777	143849	53,5	-0,176	0,077	-	-
trnV-GAC	+	72	69018	69089	48,6	-0,081	0,200	-	-
trnV-GAC_2	-	72	105354	105425	48,6	-0,081	0,200	-	-
trnV-UAC	+	654	138197	138850	33,8	0,048	0,023	-	-
trnW-CCA	-	74	35195	35268	52,7	-0,029	0,026	-	-

trnY-GUA	+	84	1280	1363	54,8	-0,053	0,130	-	-
trnfm-CAU	+	74	153103	153176	54,0	0,000	0,000	-	-

* Duplicated genes were indicated by "_2".

Figures

Fig. 1

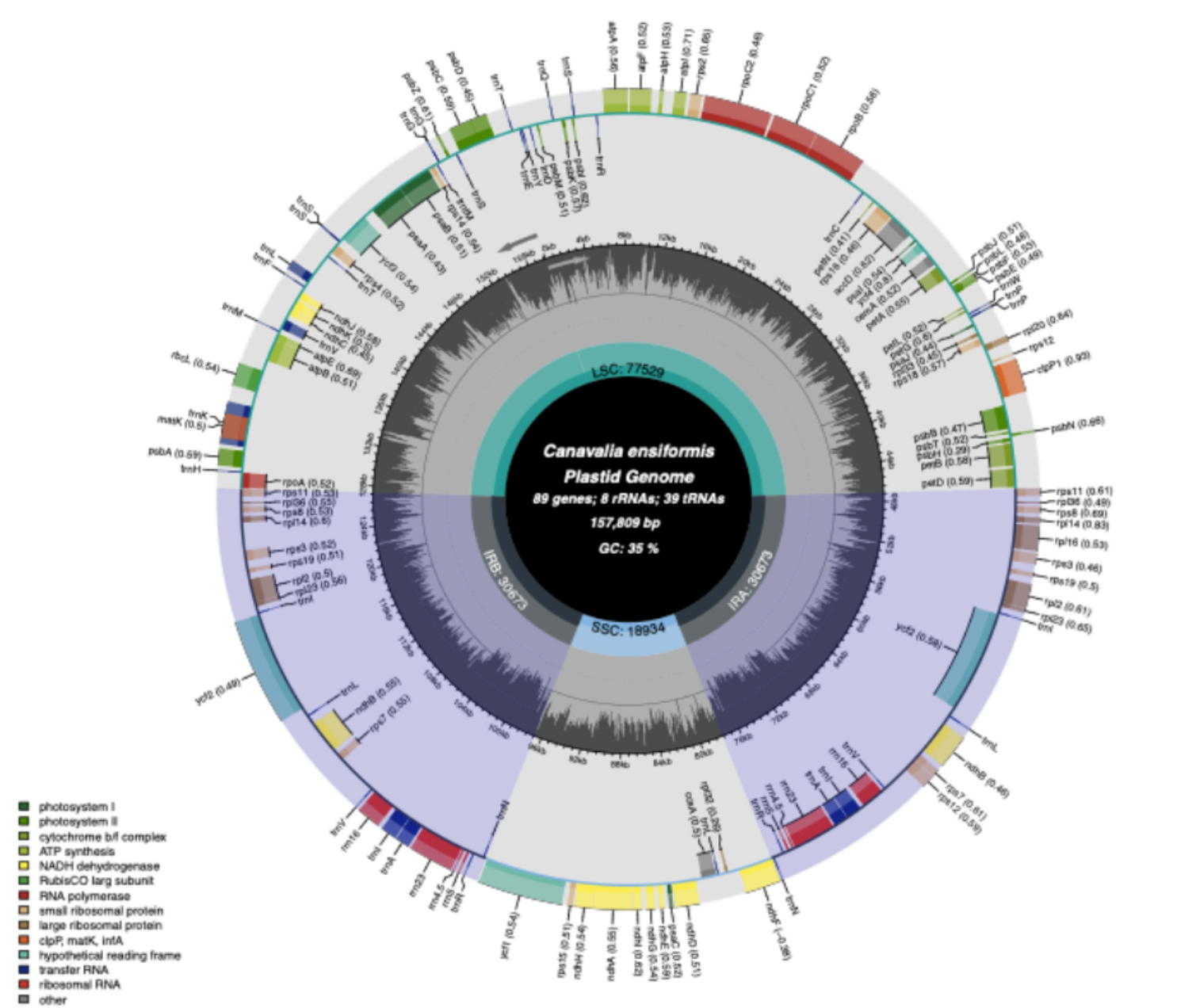


Figure 1

Fig. 2



Figure 2

Codon usage patterns of jack bean (*Canavalia ensiformis*) plastid genome. The y-axis represents the relative synonymous codon usage, whereas the x-axis represents the codons.

Fig 3.

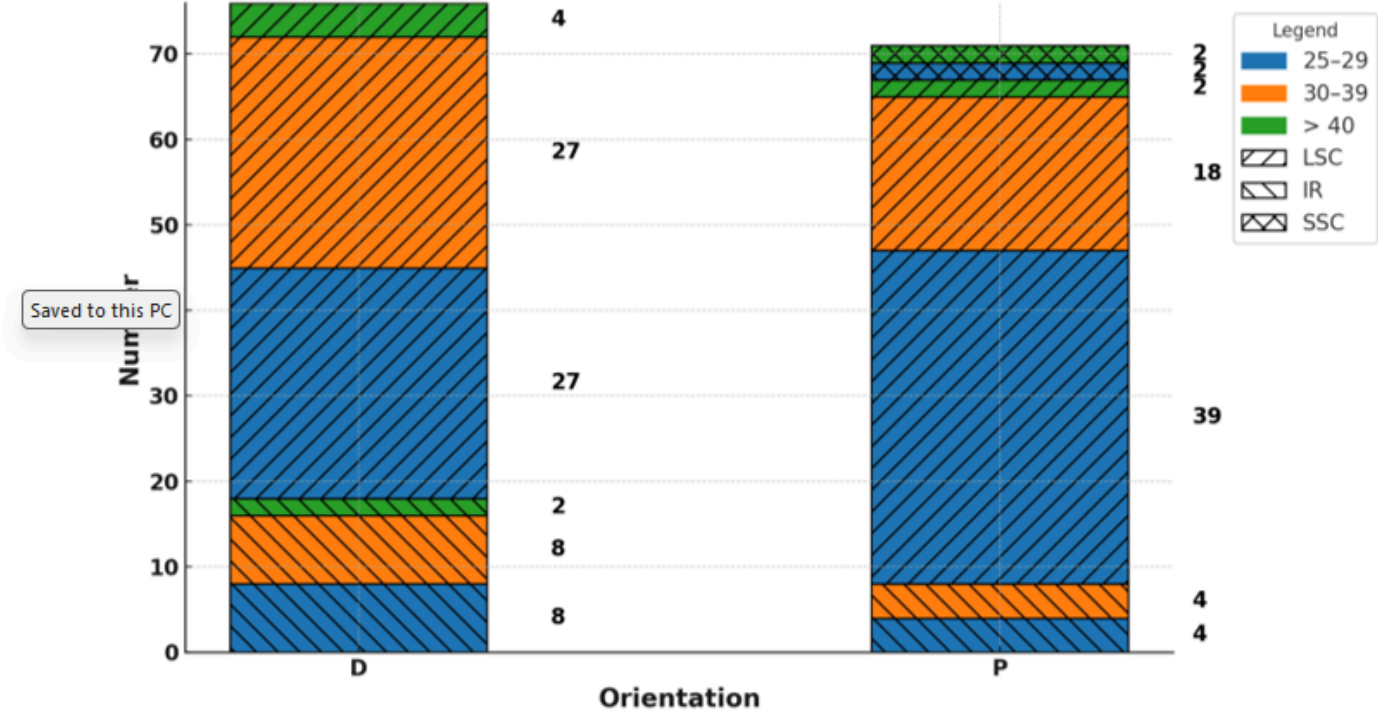


Figure 3

The number of long repeats in jack bean (*Canavalia ensiformis*). The number of different length ranges of repeats on genomic region, including long single-copy (LSC), inverted repeat (IR), and short single-copy (SSC) regions were visualized. The letters “D” stand for direct repeat type and “P” for palindromic repeat type.

Fig 4.

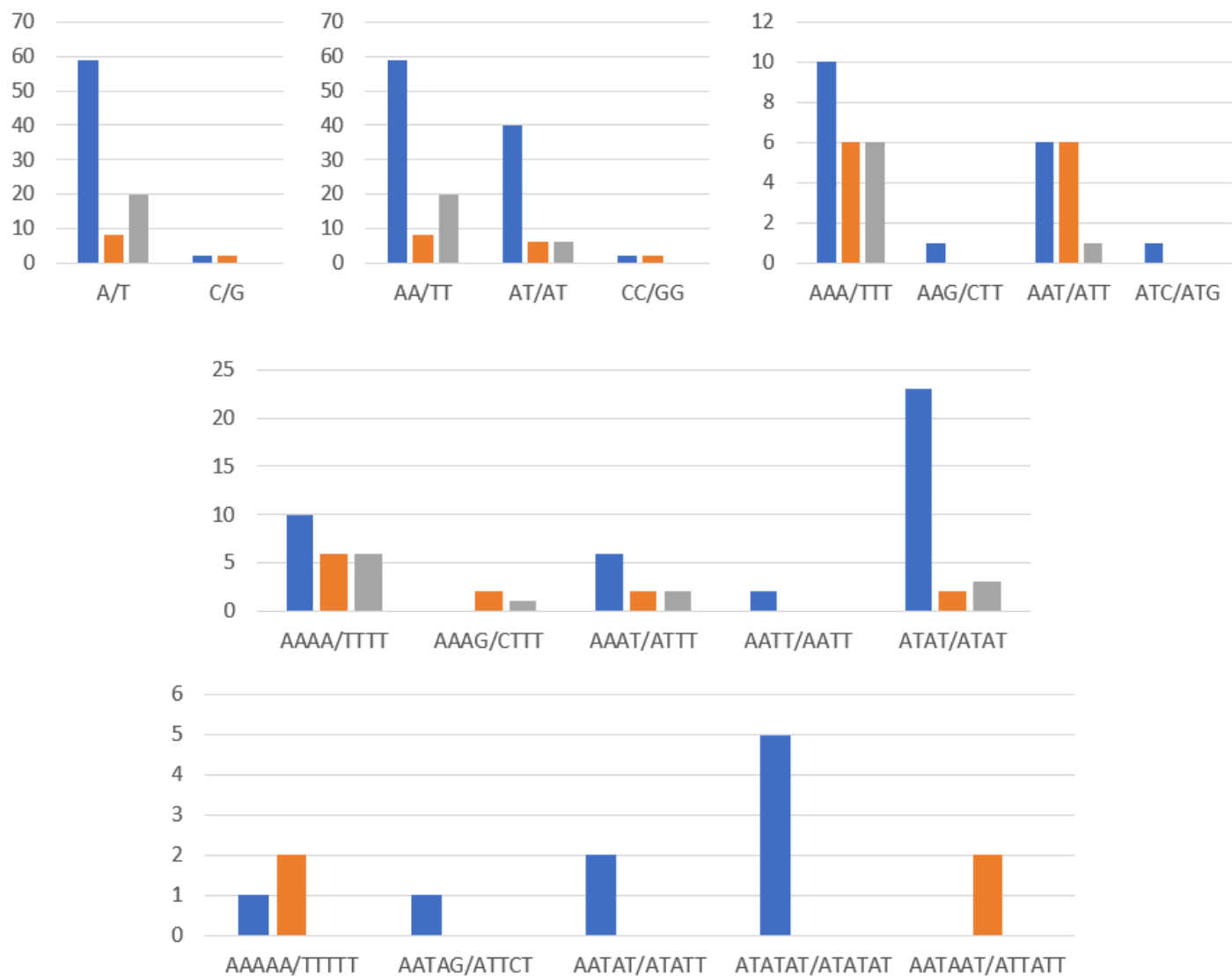


Figure 4

Number and distribution of SSR motifs in jack bean (*Canavalia ensiformis*) plastid genomes. The bar chart shows different SSR motif types in long single-copy (LSC), inverted repeat (IR), and short single-copy (SSC) regions.

Fig 5.

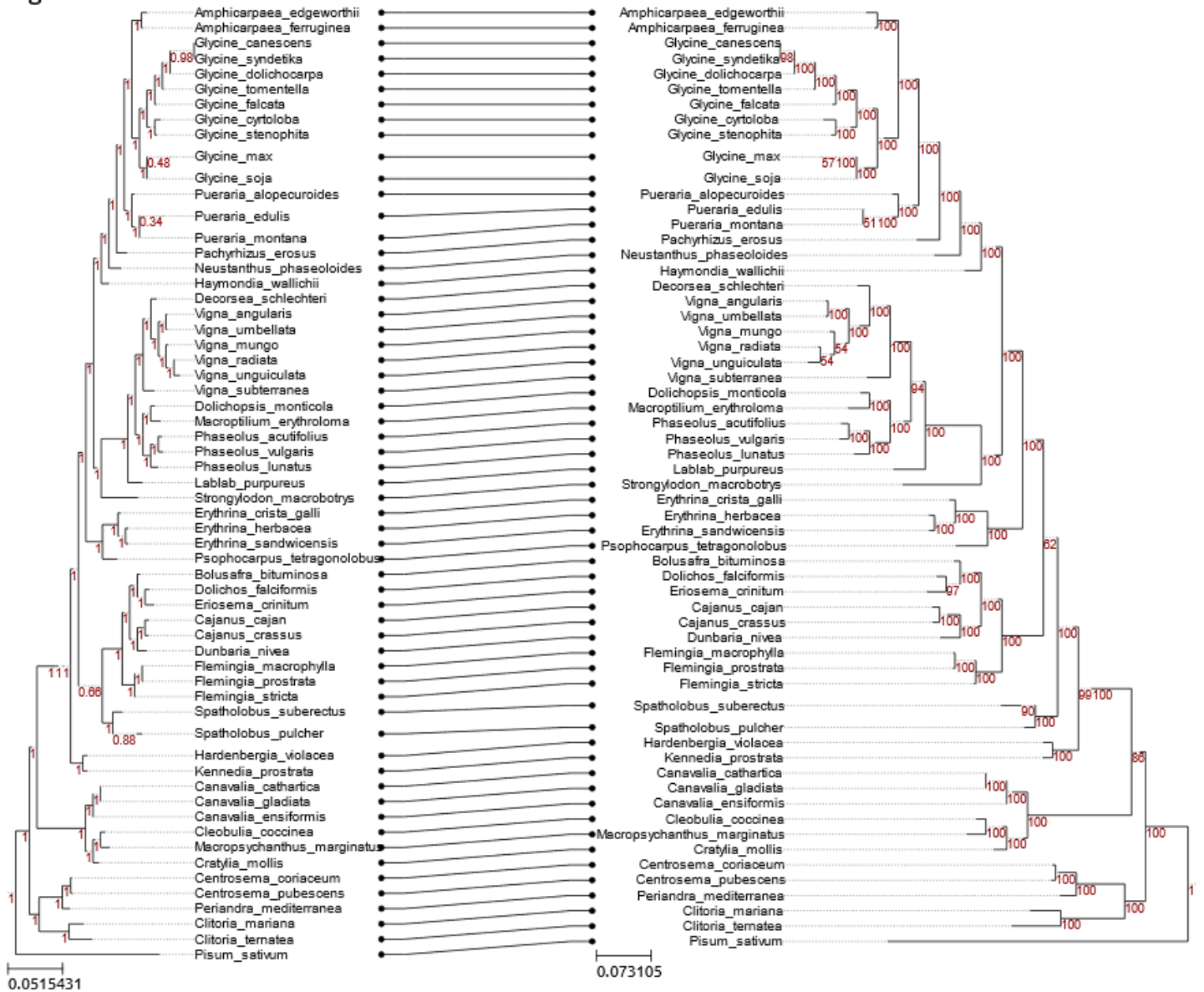


Figure 5

Phylogenetic tree of 60 species of Phaseoleae (Fabaceae) based on complete plastome sequences by maximum likelihood (ML) and Bayesian inference (BI) with *Pisum sativum* as an outgroup. Numbers at each node are the maximum-likelihood bootstrap support/Bayesian posterior probabilities.

Fig 6.

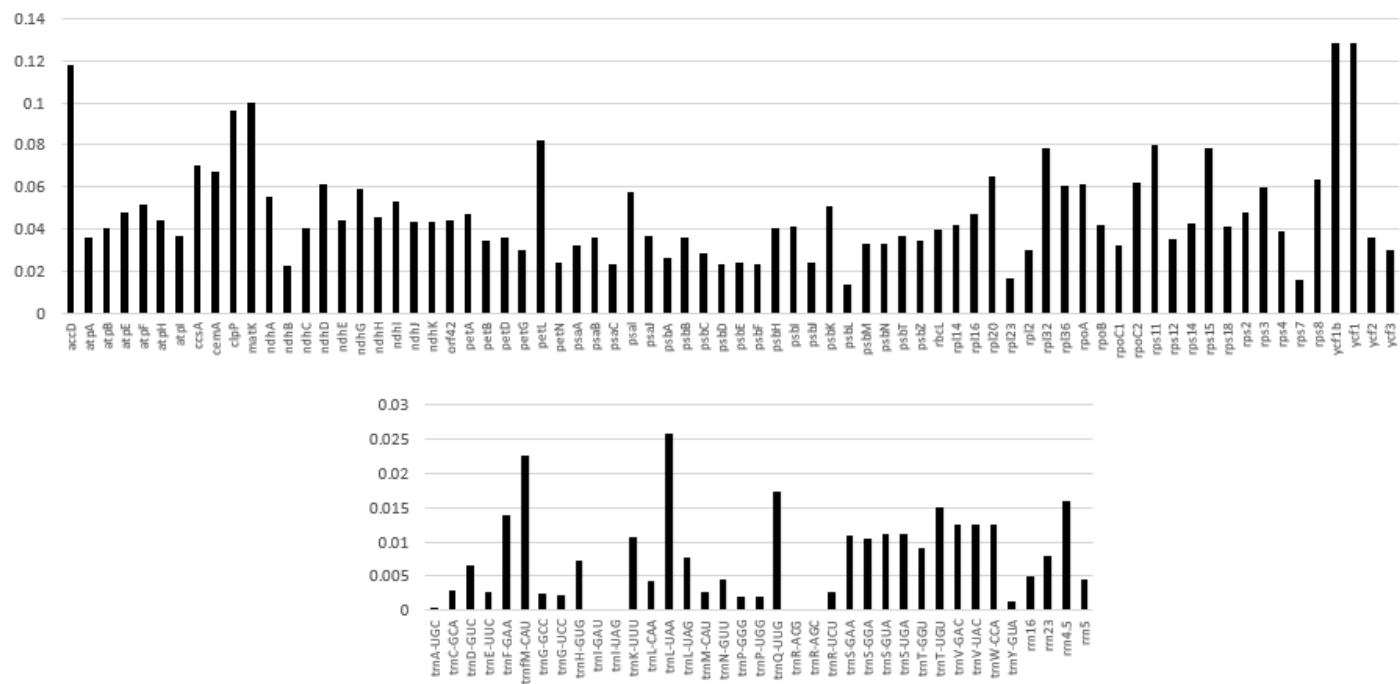


Figure 6

Comparison of nucleotide diversity (Pi) values among the Phaseoleae (Fabaceae) species. X and Y axes indicate the position of the mid-point and Pi of each window, respectively.