Chloroplast genome sequencing in winged bean (Psophocarpus tetragonolobus L.) and comparative analysis with other legumes

Nikhil Kumar Singh
ICAR-Indian Institute of Agricultural Biotechnology
Binay K. Singh
ICAR-Indian Institute of Agricultural Biotechnology
Anupama Giddhi
ICAR-Indian Institute of Agricultural Biotechnology
Harsha Srivastava
ICAR-National Institute for Plant Biotechnology
Avinash Pandey
ICAR-Indian Institute of Agricultural Biotechnology
Sudhir Kumar
ICAR-Indian Institute of Agricultural Biotechnology
A. Pattanayak
ICAR-Indian Institute of Agricultural Biotechnology
V. P. Bhadana
ICAR-Indian Institute of Agricultural Biotechnology
Sujay Rakshit
ICAR-Indian Institute of Agricultural Biotechnology
Kishor U. Tribhuvan
kishor.tribhuvan@icar.gov.in

ICAR-Indian Institute of Agricultural Biotechnology

Article

Keywords: Winged bean, Plastome, SSR, AKWB-1, Phylogenetic tree, Codon Usage Bias

Posted Date: July 19th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4615004/v1

Additional Declarations: No competing interests reported.

Abstract

The winged bean (*Psophocarpus tetragonolobus*) is a fast-growing, underutilized legume thriving in hot, humid regions. It forms symbiotic associations with a broad-spectrum cowpea rhizobial group, making it ideal for crop rotation or intercropping systems. Winged bean seeds are rich in protein, fiber, vitamins, minerals, fat, and carbohydrates, highlighting its potential as a valuable agricultural crop. In this study, we conducted whole-genome sequencing of the winged bean chloroplast using high-coverage short-read sequencing on the Illumina platform, generating over 1 billion paired-end raw reads. We utilized the GetOrganelle toolkit to assemble the chloroplast genome comprising 130 genes, including 85 protein-coding genes, 37 tRNAs, and eight rRNA genes. We also identified 84 perfect SSRs, two compound SSRs, and 15 VNTRs. Our analysis revealed the typical quadripartite structure of the chloroplast genome, along with insights into its functional classification and phylogenetic relationships with other legumes. Additionally, we identified possible genomic rearrangements through synteny analysis. Characterizing the winged bean chloroplast genome provides crucial resources for research and crop improvement. Comparative genomics of the chloroplast offers significant insights into the evolutionary and molecular biology of legumes.

Introduction

Winged bean (*Psophocarpus tetragonolobus* (L.) DC.) (2n = 2× = 18; 782 Mbp) is an underutilised legume of the family Leguminosae [1, 2]. It grows in hot, humid, equatorial countries of Southern Asia, Melanesia, and the Pacific area [3]. It is a short-day self-pollinated species but can experience cross-pollination up to 16.0% [4]. This tropical legume, maturing in 4–5 months, yields up to 14 quintals of dry seeds and 115 quintal tubers per hectare, even with minimal management [5]. Its rapid growth and vining nature make it suitable for use as a cover crop [6, 7]. Moreover, it is highly suited for crop rotation or intercropping systems as it forms symbiotic associations with a broad-spectrum cowpea rhizobial group [8] .

Winged bean is often referred to as the 'one-stalk supermarket' due to its versatile culinary applications worldwide. Various plant parts, including leaves, branches, flowers, seeds, fruits, and tubers, are utilised [3, 9]. Often dubbed the "soybean of the tropics," winged bean seeds offer crude protein (~ 34.0%), similar to soybean (~ 35.0%) [10]. Additionally, it is rich in fiber, vitamins, minerals, and carbohydrates. Winged bean seeds contain 15–18% fat, constituted by 30–40% saturated and 60–70% unsaturated fatty acids [11]. Because of its high oxidative stability, solid fat content, and good thermal conductivity, winged bean seed oil is considered superior to soybean oil [12]. Winged bean seed powder is a valuable flour that can be brewed to make a coffee-like drink [13]. The winged bean tubers have ivory flesh and contain 12–19% protein and 1–4% fat [14].

The origin of winged bean is subject to debate with two primary hypotheses. The first hypothesis suggests an African origin, proposing either *in situ* domestication followed by migration to the east or trans-domestication from an African progenitor species later in the Indian Ocean rim of Asia [2, 3]. The second hypothesis posits that winged bean is distinct from current African members of the genus and arose through allopatric speciation preceding any domestication processes [15]. However, the precise origin and progenitor of winged bean remain unresolved.

Chloroplasts originated from a cyanobacterium through endosymbiosis with a eukaryotic host around a billion years ago, creating an autotrophic line of nucleus-containing cells [16]. Throughout evolution, they retained

essential genes for replication, transcription, photosynthesis, and several other critical metabolic pathways while losing much of their original genome. Chloroplast genomes are mostly maternally inherited in angiosperms and paternally inherited in gymnosperms [17]. Angiosperm chloroplast genomes typically have a quadripartite structure, ranging in size from 120 kb to 200 kb, including a double-stranded closed loop with a long single-copy sequence (LSC, 80 kb-90 kb), a short single-copy sequence (SSC, 16 kb-27 kb), and two inverted repeats (IRs, 20 kb-28 kb) with roughly equal length. The IRs divide the chloroplast genome into LSC and SSC regions [18, 19].

Chloroplast genomes exhibit relative stability and conservation across species, with a lower mutation rate compared to the nuclear genome [20]. Nevertheless, diversification occurs in chloroplast genomes, resulting in variations in size and organisation. The expansion/contraction or loss of IRs and gene loss/duplication outside the IR are major factors contributing to size variation [21]. Loss of IRs leads to a dynamic arrangement of the chloroplast genome [22, 23]. Within the Leguminosae family, chloroplast genomes have undergone extensive rearrangements, with some legume species experiencing complete loss of one copy of the IR [24, 25]. The Genistoids, Dalbergioids, and the Old World clades of the Leguminosae species displayed several inversions, particularly a 50 kb inversion in the LSC region [26]. Additionally, during plant evolution, chloroplast genes have been lost, with some transferred to the nucleus, such as *rpl22*, *infA*, and *accD* genes [27, 28, 29]. The loss of introns from *rps12* and *clpP* is also reported in the legume genome [22, 27].

Chloroplast genomes have been extensively used in evolution, phylogeny, and phylogeography studies [30, 31, 32]. In the Leguminosae family, comprising approximately 751 genera and 19,500 species ranging from trees to herbaceous crop plants [33], chloroplast-derived markers, particularly *matK* gene and the *trnL-trnF* intergenic spacer have played a crucial role in exploring evolutionary relationships [34, 35, 36]. Chloroplast genomic data have also been extensively employed to study gene expression and regulation, including RNA editing sites and codon usage bias [37, 38, 39]. Besides, simple sequence repeats (SSRs) within chloroplast genomes are potential DNA markers for species identification [40] .

Chloroplast genetic engineering offers unique advantages such as high-level transgene expression [41], multigene engineering in a single transformation event [42], transgene containment *via* maternal inheritance [43], lack of gene silencing [44], absence of position effect [45], pleiotropic effects [46], and prevention of undesirable foreign DNA [47]. Complete chloroplast genome sequences are essential for identifying spacer regions for transgene integration at optimal sites through homologous recombination, as well as for determining endogenous regulatory sequences for optimal transgene expression [48].

Traditional methods for obtaining chloroplast genome sequences involve chloroplast DNA isolation, random shearing, cloning into large-insert size vectors, and shotgun sequencing. Recent advancements, such as whole-genome PCR amplification with universal primers and high-throughput sequencing, have introduced faster and cost-effective approaches [49, 50, 51]. Next Generation Sequencing (NGS) technology, particularly platforms like Illumina, has significantly accelerated chloroplast genome sequencing. Moore et al.[52] made the pioneering attempt to use NGS for chloroplast genome sequencing, leading to the sequencing of numerous chloroplast genomes, with Illumina being the most commonly used platform [53, 54, 55, 56]. Despite over 44 published chloroplast genomes within the Leguminosae family, the chloroplast genome of winged bean remains unexplored. In this study, genomic DNA from fresh young leaves of winged bean was sequenced using the Illumina HiSeq2500 platform, and the chloroplast genome was assembled using the embryophyta plastid database as reference. The

study elucidated the winged bean chloroplast genome sequence and its characteristics and compared it with those of other Leguminosae species. This research aims to enhance our understanding of the winged bean chloroplast genome and provide valuable markers for phylogenetic and genetic studies.

Materials and Methods

Plant material

The study utilised the dual-purpose cultivar AKWB-1 of winged bean, which serves both as a vegetable and a pulse. Given the significant level of cross-pollination reported in winged bean, the AKWB-1 plants were selfed for three successive generations to achieve homozygosity. The winged bean plants were grown at the experimental farm of the ICAR-Indian Institute of Agricultural Biotechnology, located in Ranchi, Jharkhand, India, with the geographical coordinates 23°16'27.6"N, 85°20'29.4"E.

DNA isolation, library preparation, sequencing, and annotation

The young leaves of winged bean were utilised to extract high molecular weight genomic DNA using the Cetyl Trimethyl Ammonium Bromide (CTAB) method [57]. To remove RNA contamination from the DNA, 2.0 µl of RNase A (10 mg/ml, HiMedia) was added to 20 µl of DNA dissolved in TE buffer (Tris–EDTA, pH = 8.0), followed by incubation at 37°C for 3–4 h. The quality of the purified DNA samples was analysed on a 1.0% agarose gel and quantified based on the absorbance at 260 nm using a NanoDrop™ OneC microvolume UV–Vis spectrophotometer (Thermo Scientific). The whole-genome sequencing library was constructed using the TruSeq DNA PCR-Free Library Prep Kit (Illumina, catalog no. 20015963) following the manufacturer's instructions. The prepared library was quantified using a Qubit 3 fluorometer (Thermo Fisher Scientific, USA) with the DNA High Sensitivity kit, and the library size was verified using a Bioanalyzer 2100 (Agilent Technologies, CA, USA). After assessing the quality and quantity of the library, whole-genome sequencing was conducted on a Illumina HiSeq2500 instrument with paired-end (2 × 150 bp) sequencing strategy using the TruSeq Rapid SBS Kit (Illumina, catalog no. FC-402-4023). FastQC v0.11.9 software (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was employed to evaluate the quality of sequencing data, including base qualities, GC content, adapter content, and overrepresentation analysis. Adapter sequences were trimmed using fastp v0.20.1 software [58] (https://github.com/OpenGene/fastp), with a minimum length of two bases, quality filtering disabled, and forced poly-G trimming. The high-quality reads were *de novo* assembled into the chloroplast genome using the GetOrganelle v1.5.1c toolkit [59], specifying the embryophyta plastid database. Annotation of the complete chloroplast genome was performed with GeSeq and manual corrections [60]. The complete chloroplast genome sequence of the winged bean was submitted to GenBank with the accession number PP894786.1.

Phylogenetic Analysis

The nucleotide sequences of all predicted chloroplast genes of *P. tetragonolobus* and the reported chloroplast genes for nine other related species, including *Vigna radiata*, *Phaseolus vulgaris*, *Lablab purpureus*, *Cyamopsis tetragonoloba*, *Glycine max*, *Cajanus cajan*, *Medicago truncatula*, *Cicer arietinum*, and *Arachis hypogaea*, along with *Arabidopsis thaliana* as the outgroup, were obtained from NCBI. Alignment and subsequent phylogenetic analysis were conducted using the genes shared across the species. The sequences were aligned using MAFFT v7 [61] with 1000 iterative refinement steps using "--maxiterate 1000". The resulting aligned sequences were saved

in FASTA format. Phylogenetic trees for each aligned gene were inferred using RAxML-NG [62] with 1000 Bootstrap replicates and the "General Time Reversible (GTR) with gamma-distributed rate variation among sites (GTR + G)" model. The best genetree files were used to prepare a multigene-based species tree with ASTRAL v5.7.8 [63]. Finally, the phylogenetic tree was visualised using FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

Synteny and genome rearrangement analysis

We downloaded the chloroplast genomes of *G. max*, *C. cajan*, and *L. purpureus*, along with *A. thaliana* as an outgroup species from CpGDB (https://www.gndu.ac.in/CpGDB/) and analysed the synteny and rearrangement in the chloroplast genomes of *P. tetragonolobus* and related legumes using two different methods. For synteny plots, pairwise BLASTN results were generated using the whole chloroplast genome of each species in R using a custom script. Information on BLAST hits was visualised in R using the genoPlotR package [64]. Subsequently, the alignments were annotated using species-specific GFF3 specifications. Information on BLAST hits among homologous segments was also visualised in R using the genoplotR package [64]. As a complementary approach, we performed global alignment of the chloroplast genomes with *P. tetragonolobus* as the reference. The alignment was conducted on the online platform mVISTA in Shuffle-LAGAN mode [65].

Estimation of adaptive evolution of protein-coding genes

To analyse the degree of non-synonymous (Ka) and synonymous (Ks) substitutions, as well as their ratio (dN/dS), the coding DNA sequences (CDS) of *P. tetragonolobus* were compared with *G. max*, *C. cajan*, and *L. purpureus*, along with *A. thaliana.* For this, we performed pairwise alignments using MAFFT v7 [66], and then Ka, Ks substitutions, and their ratio were calculated using KaKs_Calculator 3 [67] with the MA model.

Codon usage analysis

The codon usage within the coding part of the chloroplast was calculated using CodonW v1.4.4 available at https://codonw.sourceforge.net using universal codon standards. RSCU (relative synonymous codon usage) values were plotted using R package ggplot [68].

Repeat Analysis

The long repeats in the winged bean chloroplast genome was analysed using REPuter [69]. Repeats were reported as either F (forward), R (reverse), P (palindromic), or C (complement), with parameters set at a Hamming Distance of 3 and a Minimal Repeat Size of 30. The chloroplast genome was also screened for perfect SSRs (pSSRs), compound SSRs (cSSRs), and variable number tandem repeats (VNTRs) using Krait v1.3.3 software [70]. The screening was conducted based on specific criteria: mono-nucleotide repeat motifs were required to have a minimum of 10 repeats, di-nucleotide repeat motifs required a minimum of five repeats, tri-nucleotide repeat motifs needed at least five repeats, while tetra-, penta-, and hexa-nucleotide repeat motifs were required to have a minimum of four repeats. If the distance between two SSRs was < 10 bp, they were considered as cSSR.

Results

Assembly and annotation of winged bean chloroplast genome

We used the Illumina HiSeq2500 next-generation sequencing platform for whole-genome sequencing of winged bean, generating 1,030,974,930 paired-end raw reads. After cleaning for adaptors and low-quality reads,

1,030,150,624 clean reads were obtained. Subsequently, the GetOrganelle v1.5.1c toolkit was employed to assemble the chloroplast genome from the winged bean whole-genome sequencing data. Using a modified baiting and iterative mapping approach, GetOrganelle successfully recruited 36,123,587 chloroplast-associated clean reads, constituting 3.51% of the total clean reads. This approach facilitated the assembly of the complete winged bean chloroplast genome, totaling 151,571 bp, with > 35,000X coverage (**Supplementary Table S1**).

The chloroplast genome of winged bean exhibited a typical quadripartite structure, consisting of an LSC region of 82,736 bp, an SSC region of 17,777 bp, and a pair of equal-sized IRs each measuring 25,529 bp. The winged bean chloroplast genome had an overall GC content of 35.26%, with the LSC, SSC, and IR regions showing GC contents of 32.63%, 28.55%, and 41.86%, respectively (Table 1).

Table 1
Characteristics of the winged bean chloroplast genome.

Attribute	Start	End	Size (bp)	Adenine (A)	Thymine (T)	Guanine (G)	Cytosine (C)	AT (%)	GC (%)
Whole chloroplast genome (bp)	1	151571	151571	49135 (32.42%)	48995 (32.32%)	26888 (17.74%)	26553 (17.52%)	64.74	35.26
Large single-copy region (bp)	68836	151571	82736	27979 (33.82%)	27763 (33.56%)	13833 (16.72%)	13159 (15.90%)	67.37	32.63
Small single-copy region (bp)	25530	43306	17777	6313 (35.51%)	6388 (35.93%)	2368 (13.32%)	2707 (15.23%)	71.45	28.55
Inverted repeat B (IRb) (bp)	1	25529	25529	7379 (28.90%)	7462 (29.23%)	5547 (21.72%)	5140 (20.13%)	58.14	41.86
Inverted repeat A (IRa) (bp)	43307	68835	25529	7462 (29.23%)	7380 (28.90%)	5140 (20.13%)	5547 (21.73%)	58.14	41.86
Prt. coding genes	-	-	78004	24940 (31.97%)	24929 (31.95%)	14292 (18.32%)	13843 (17.75%)	63.93	36.07
tRNA	-	-	2746	654 (23.82%)	639 (23.27%)	718 (26.15%)	735 (26.77%)	47.09	53.52
rRNA	-	-	8667	1951 (22.51%)	1952 (22.52%)	2381 (27.47%)	2383 (27.49%)	45.03	54.96

The comparison of the chloroplast genome of winged bean with the chloroplast genomes of related legume species, including *V. radiata*, *P. vulgaris*, *L. purpureus*, *C. tetragonoloba*, *G. max*, *C. cajan*, *M. truncatula*, *C. arietinum*, and *A. hypogaea*, along with *A. thaliana* indicated that although the sizes of the overall genome had differences, the GC content was similar in LSC, SSC, and IR regions of different species. We observed a little difference in total genes, CDS and tRNAs among the ten legume species. *C. tetragonoloba* exhibited the

maximum number of protein coding genes, CDS and tRNAs and *C. arietinum* showed the least (**Supplementary Table S2**).

The chloroplast genome of winged bean consisted of 130 genes, including 85 protein-coding genes, 37 tRNAs, and 8 rRNAs genes. Individually, LSC contained 83 genes, including 61 protein-coding and 22 tRNAs genes while SSC contained 13 genes, including 12 protein-coding and one tRNA genes. The IR regions contained 17 duplicated copies of 6 protein-coding genes, 7 tRNAs, and 4 rRNAs genes **(**Fig. 1). Overall, 24 intron-containing genes (14 protein-coding genes, 8 tRNA genes, and 2 rRNA genes) were found. Among these, 22 genes had one intron, and *clpP* and *pafI* had two introns each. *trnK-UUU* had the largest intron (2585 bp) and *rrn23* had the smallest intron (198 bp) (**Supplementary Table S3**).

The functional classification of winged bean chloroplast genes indicated that 47 genes had photosynthesis-related functions, including genes for photosystem I, photosystem II, Cytochrome b/f complex, ATP synthase, NADH dehydrogenase, Rubisco large subunit, and Photosystem assembly stability factor. Similarly, 75 genes were categorised for transcription and translational related functions, which included RNA polymerases, ribosomal protein small subunit, ribosomal protein large subunit, tRNA, and rRNAs. The remaining 8 genes, including maturase RNA processing, protease gene, c-type cytochrome synthesis gene, subunit of acetyl-CoA-carboxylase (fatty acid synthesis) gene, envelope membrane protein (carbon metabolism), and hypothetical chloroplast reading frames, displayed other functions (Table 2).

Table 2
Chloroplast encoding genes in *Psophocarpus tetragonolobus*

Category	Group	Genes
Photosynthesis	Photosystem I	*psaA, psaB, psaC, psaI, psaJ*
	Photosystem II	*psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbT, psbZ*
	Cytochrome b/f complex	*atpA, atpB, atpE, atpF, atpH, atpI*
	ATP synthase	*petA, petB, petD, petL, petG, petN*
	NADH dehydrogenase	*ndhA, ndhB*[#], *ndhB*[#], *ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*
	Rubisco large subunit	*rbcL*
	Photosystem assembly stability factor	*pafI, pafII, pbf1*
Transcription and Translation	RNA polymerase	*rpoA, rpoB, rpoC1, rpoC2*
	Ribosomal protein (SSU)	*rps7, rps12*[#], *rps15, rps12*[#], *rps7, rps4, rps14, rps2, rps16, rps18, rps12, rps11, rps8, rps3, rps22, rps19*
	Ribosomal protein (LSU)	*rpl2*[#], *rpl23*[#], *rpl32, rpl23*[#], *rpl2*[#], *rpl33, rpl20, rpl36, rpl14, rpl16*
	Transfer RNA	*trnI-CAU*[#], *trnL-CAA*[#], *trnV-GAC*[#], *trnI_GAU*[#], *trnA-UGC*[#], *trnR-ACG*[#], *trnN-GUU*[#], *trnL-UAG, trnN-GUU*[#], *trnR-ACG*[#], *trnA-UGC*[#], *trnI-GAU, trnV-GAC*[#], *trnL-CAA*[#], *trnI-CAU*[#], *trnH-GUG, trnK-UUU, trnM-CAU, trnV-UAC, trnF-GAA, trnL-UAA, trnT-UGU, trnS-GGA, TrnfM-CAU, trnG-GCC, trnS-UGA, trnT-GGU, trnE-UUC, trnY-GUA, trnD-GUC, trnC-GCA, trnR-UCU, trnG-UCC, trnS-GCU, trnQ-UUG, trnW-CCA, trnP-UGG*
	Ribosomal RNAs	*rrn16*[#], *rrn23*[#], *rrn4.5*[#], *rrn5*[#], *rrn5*[#], *rrn4.5*[#], *rrn23*[#], *rrn16*[#]
Other genes	Maturase RNA processing	*matK*
	Protease	*clpP1*
	c-type cytochrome synthesis gene	*ccsA*
	Subunit of acetyl-CoA-carboxylase (fatty acid synthesis)	*accD*
	Envelope membrane protein (carbon metabolism)	*cemA*

[#] The present in the IR region has duplicate copies

Category	Group	Genes
	hypothetical chloroplast reading frames	*ycf1, ycf2, ycf2*

[#] The present in the IR region has duplicate copies

Phylogenetic relationships with other legumes

The phylogenetic position of *P. tetragonolobus* within the Leguminosae family was determined by aligning chloroplast gene sequences with related species, including *V. radiata*, *P. vulgaris*, *L. purpureus*, *C. tetragonoloba*, *G. max*, *C. cajan*, *M. truncatula*, *C. arietinum*, and *A. hypogaea*, with *A. thaliana* serving as the outgroup. The resulting multigene-based phylogenetic tree showed robust support (bootstrap value 1.0) for all resolved nodes. The analysis revealed two major clades among the studied legume species. *A. hypogaea* was affiliated with the Dalbergioid clade, while the remaining species formed an Old World clade. *C. tetragonoloba*, belonging to the Indigoferoid group (Tribe Indigofereae), clustered closely with *C. cajan*, *P. tetragonolobus*, *G. max*, *L. purpureus*, *P. vulgaris*, and *V. radiata*, all of which are part of the Millettioid group under Tribe Phaseoleae within the Phaseoloid clade. Additionally, the Old World clade included *C. arietinum* (Tribe Cicereae) and *M. truncatula* (Tribe Trifolieae), which clustered together as part of the Inverted Repeat Lacking Clade (IRLC) within the major Hologalegina clade (Fig. 2).

Analysis of synteny and genomic rearrangements

To identify possible occurrences of rearrangements in the chloroplast genomes, we analysed the synteny of whole chloroplast genome sequences of *C. cajan*, *G. max*, and *L. purpureus* along with *A. thaliana* as an outgroup species (Fig. 3). We used two approaches: a) pairwise and b) global alignment. The pairwise alignments presented high synteny but a strong signature of rearrangements and inversions involving the IR and SSC regions. Specifically, the genomic segment of the SSC region appeared to be inverted in *P. tetragonolobus* compared to *C. cajan* and *G. max*. The IR region flanking the SSC showed high synteny with other legumes. Since the chloroplast is a circular genome, the visible rearrangement was also confirmed with gene order analysis using *P. tetragonolobus* as a reference.

The global alignment, using the shuffle-LAGAN algorithm incorporated in the m-Vista pipeline and winged bean chloroplast genome annotations, indicated that within the region between 24.5 to 46 kb, the exons were comparatively less conserved than the rest of the chloroplast (Fig. 4). This was the same region where the inversion appeared to occur in the pairwise synteny analysis.

Analysing selective pressure on protein-coding genes

To assess selective pressures acting on protein-coding genes in the chloroplast genome, the Ka/Ks ratio was calculated among *P. tetragonolobus*, *C. cajan*, *G. max*, and *L. purpureus*, along with *A. thaliana* (Fig. 5). The Ka/Ks ratios ranged between 0.001 to 1.4. The highest and only gene with a ratio > 1 was *rpl23* in pairwise comparison to *A. thaliana*. Similarly, *rpl23* and an additional gene, *rpl2*, exhibited Ka/Ks values of 0.87 and 0.96, respectively, in comparison to *C. cajan*, suggesting significant evolutionary divergence. When compared with *L. purpureus*, the only gene with a Ka/Ks value close to 1.0 was ndhB, with a value of 0.8. The remaining genes had Ka/Ks values ≤ 0.5.

Comparative analysis of codon usage frequency

Sixty-one distinct codons for 20 different amino acids were identified in the chloroplast genome of winged bean (Fig. 6). The termination codons were excluded from the analysis. Among the 61 distinct codons, 6 individual codons encoded Arginine, Leucine and Serine; 4 codons encoded Alanine, Glycine, Proline, Threonine, and Valine; and 3 codons encoded Isoleucine; and 2 codons encoded Asparagine, Aspartic acid, Cysteine, Glutamic acid, Glutamine, Histidine, Lysine, Phenylalanine, and Tyrosine. Methionine and Tryptophan were encoded by 1 codon each. The RSCU values ranged between 0.47 and 4.38, with CGC coding for alanine having the lowest RSCU, while TTT for phenylalanine had the highest. Phenylalanine, lysine, and asparagine were the most abundant amino acids coded by the winged bean genome.

Long-repeats and SSRs

The analysis of long repeats in the winged bean chloroplast genome showed 38 palindromic, 16 forward, three reverse and two complement repeats, thus in total 59 long repeats (Fig. 7b). Among them, 38 repeats were in the 30–35 bp range while 5 and 7 repeats were in the 36–40 bp and 41–45 bp range, respectively. Nine repeats were found to be more than 45 bp in length (Fig. 7a**).** The longest repeat was of length 287 bp. Among the 59 long repeats, LSC contained 35 repeats while SSC and IRs contained 6 and 18 repeats, respectively. In total 68% (n = 40) of the 59 repeats were present in the intergenic spacers while the rest 32% (n = 19) were found to be overlapping with gene *ycf2, ndhA, ndhF, rps12, rpl22, pafI and psaA* (**Supplementary Table S4**).

We identified 84 perfect SSRs, two compound SSRs, and 15 VNTRs (Fig. 7c; **Supplementary Table S5**) accounting for one perfect SSR per 1.8 kbp, one compound SSR per 75.79 kbp, and one VNTR per 10.10 kbp of the winged bean chloroplast genome. Among the 84 perfect SSR loci analysed, 49 (58.33%) were mononucleotides while the remaining 35 (41.67%) were dinucleotides (Fig. 7e). A/T repeats were the most frequent mononucleotide repeats (43, 51.19%) followed by G/C (6, 7.14%). Ten bp SSRs were most frequent (53, 63.09%) followed by 11 bp (14, 16.67%), 12 bp (10, 11.90%), 14 bp (4, 4.76%), 16 bp (2, 2.38%) and 15 bp (1, 1.19%). Dinucleotide repeats contained four types of repeat motifs (AG/CT, TA, TC, and AT), of which AT accounted for 22.61% (19) followed by TA (13, 15.48%), AG/CT (2, 2.38%), and TC (1, 1.19%) (Fig. 7d). The frequency of the repeat motifs varied from 33.33% (n = 10) to 1.19% (n = 14 &15).

A total of six distinct repeat motif types were identified in the study. Different repeat motifs were reiterated 5 to 16 times (Table 3).

Table 3
Frequency distribution of perfect SSR repeat motifs in winged bean chloroplast genome.

S. No.	Repeat motif	Number of reiterations of the motif									Total
		5	6	7	10	11	12	14	15	16	
1	A/T	-	-	-	23	14	3	-	1	2	43
2	G/C	-	-	-	5	-	-	1	-	-	6
3	AG/CT	2	-	-	-	-	-	-	-	-	2
4	AT	13	3	3	-	-	-	-	-	-	19
5	TA	9	4	-	-	-	-	-	-	-	13
6	TC	1	-	-	-	-	-	-	-	-	1
	Total	25	7	3	28	14	3	1	1	**2**	**84**

Among the 84 perfect SSRs, two compound SSRs, and 15 VNTRs identified in the study, LSC contained the maximum number of perfect SSRs (57, 67.86%), compound SSRs (02, 100%), and VNTRs (09, 60%) followed by SSC with 18 (21.43%) perfect SSRs and 03 (20%) VNTRs, IRs with 09 (10.71%) perfect SSRs and 03 (20%) VNTRs. Regarding the distribution of SSRs and VNTRs in the genic and non-genic regions, the maximum number of perfect SSRs (57, 67.86%), compound SSRs (02, 100%), and VNTRs (10, 66.67%) were located in the intergenic spacers followed by exon with 16 (19.05%) perfect SSRs and 04 (26.67%) VNTRs. The introns accounted for the lowest number with 11 (13.10%) perfect SSRs and 01 (6.67%) VNTRs. The perfect SSR containing genes were *rpl2*, *ndhA*, *ycf1*, *rpl2*, *trnK-UUU*, *atpB*, *pafI*, *rpoB*, *rpoC1*, *rpoC2*, *rps2*, *atpF*, and *rps18*. *ycf1* contained 11 perfect SSRs. The VNTR containing genes were *ycf2*, *rps7*, *ccsA*, and *psbB* (**Supplementary Table S5**).

Discussion

The legume family (Leguminosae) is economically one of the most successful lineages among flowering plants [33]. It exhibits a significantly higher species diversification rate over the last 60 million years compared to angiosperms as a whole [71]. Within this family, winged bean distinguishes itself with its nutrient-rich components and effective symbiotic associations with a broad spectrum of rhizobia strains, making it suitable for low-input and self-resilient agricultural systems [72]. Chloroplast genomes, with their conserved nature, are crucial resources for studying evolutionary dynamics and phylogenetic relationships across plant taxa [73, 74]. The present study reports the sequencing and characterisation of the chloroplast genome of *P. tetragonolobus*, along with its comparative analysis with other legumes, including *V. radiata*, *P. vulgaris*, *L. purpureus*, *C. tetragonoloba*, *G. max*, *C. cajan*, *M. truncatula*, *C. arietinum*, and *A. hypogaea*, with *A. thaliana* serving as the outgroup.

Leguminosae is divided into three subfamilies: Caesalpinioideae, Mimosoideae, and Papilionoideae. Caesalpinioideae, a paraphyletic group, is the ancestral base for the monophyletic subfamilies Mimosoideae and Papilionoideae [75]. Papilionoideae, the largest subfamily, comprises 13,800 species across 28 tribes in 478 genera [33]. It is further divided into Swartzioid and Aldinoid lineages and other genera within a larger monophyletic group marked by a 50 kb inversion in the chloroplast genome [24]. The 50 kb inversion group includes three major clades: Genistoids, Dalbergioids, and the Old World clade. The Genistoid clade is characterised by the accumulation of

quinolizidine while the Dalbergioid clade typically exhibits "aeschynomenoid" root nodule morphology [76]. The Old World clade further segregates into the Indigoferoid/Millettioid and Hologalegina clades, with the latter splitting into the Robinioid and Inverted Repeat-Lacking Clade (IRLC). Indigofereae is sister to the Millettioid group, comprising Phaseoloid and core Millettieae clades and allies [77, 78].

In recent years, there has been a growing interest in the legume systematics community to combine expertise and data to capitalise on new approaches in genetics and bioinformatics.With the advancement of sequencing technologies, an increasing number of chloroplast genomes have been sequenced and used for phylogenetic analysis. We used the nucleotide sequences of all predicted chloroplast genes of *P. tetragonolobus* and the reported chloroplast genes for nine other legumes, including *V. radiata*, *P. vulgaris*, *L. purpureus*, *C. tetragonoloba*, *G. max*, *C. cajan*, *M. truncatula*, *C. arietinum*, and *A. hypogaea*, along with *A. thaliana* as the outgroup, to delineate the phylogenetic position of *P. tetragonolobus* in relation to the related genera. The multigene-based phylogenetic tree resolved the cladistic position of all the legumes considered for the study with robust bootstrap support. *P. tetragonolobus* clustered closely with *C. cajan*, G. max, *L. purpureus*, *P. vulgaris*, and *V. radiata* of the Millettioid group under Tribe Phaseoleae within the Phaseoloid clade. Moreover, all the other legume species considered for the study showed affiliation with their respective clades consistent with the current state of legume phylogeny [79, 80, 81], reinforcing the utility of chloroplast genome sequences in deep phylogenetic analysis.

The comparative analysis of the sequences of the chloroplast genomes of *P. tetragonolobus* and other legumes revealed clade-wise general conservation in genome size, length of IR, LSC, and SSC regions, along with their GC contents and gene content. As expected, *C. arietinum* and *M. truncatula*, belonging to the Hologalegina/ Inverted Repeat Lacking Clade (IRLC), exhibited the smallest genome size and gene contents, attributed to the presence of only a single copy of IR [82, 83, 84]. *A. hypogaea*, belonging to the Dalbergioid clade, exhibited the largest genome size, whereas the genome sizes for Indigoferoid/Millettioids, comprising *P. tetragonolobus*, *V. radiata*, *P. vulgaris*, *L. purpureus*, *C. tetragonoloba*, *G. max*, and *C. cajan*, were slightly smaller than the Dalbergioids, ranging from 151,294 bp to 152,530 bp, varying by only 1236 bp. These results indicate a notable degree of genomic homogeneity among the leguminous species under investigation. Moreover, they highlight the strength of the cladistic approach to biological classification based on the hypotheses of most recent common ancestry.

We observed a notable uniformity in GC content across the LSC, SSC, and IR regions among various legume species. Furthermore, the GC content of tRNAs and rRNAs was significantly higher than that of protein-coding genes. Notably, a proportionately higher number of GC-rich tRNAs and rRNAs in the IR regions contributed to their overall higher GC content compared to the LSC and SSC regions. These GC-rich regions ensure structural integrity and functional resilience across diverse taxa [85, 86, 87].

The synteny analysis of whole chloroplast sequences from *P. tetragonolobus*, *C. cajan*, *G. max*, and *L. purpureus*, along with *A. thaliana*, provided valuable insights into the structural variations within these genomes. Pairwise alignments revealed high synteny among the chloroplast genomes but also unveiled a notable signature of rearrangements and inversions, particularly within the IR and SSC regions. Notably, the SSC segment exhibited an inversion in *P. tetragonolobus* compared to *C. cajan* and *G. max*, indicating a structural deviation specific to this species. Complementing the pairwise analysis, global alignment using the shuffle-LAGAN algorithm highlighted a region of reduced exon conservation spanning from 24.5 to 46 kb within the chloroplast genomes. This region corresponded to the site of inversion observed in the pairwise synteny analysis, reinforcing the presence of rearrangements within this segment of the chloroplast genome. The observed rearrangements may be attributed

to flip-flop intramolecular recombination in the plastome, a mechanism proposed by Ogihara et al.[88]. While such events are rare, recent studies [89, 90] highlight their significance as evolutionary drivers, potentially conferring adaptive advantages. Identifying structural variations within chloroplast genomes, such as inversions and rearrangements, underscores the dynamic nature of plastome evolution. Understanding the mechanistic underpinnings and functional implications of these rearrangements offers valuable insights into the evolutionary trajectories of plant species within the Leguminosae family.

The Ka/Ks ratio value, which infers the rate of gene divergence between species, serves as an indicator to identify genes undergoing different selection pressures [91, 92]. In protein-coding genes, synonymous substitutions occur more frequently than non-synonymous substitutions [91, 93]. In this study, since the majority of genes had values between 0 and 0.5, this is a strong signature representing that in *P. tetragonolobus*, nonsynonymous mutations are being removed from the population at a faster rate than synonymous mutations. Thus, the genes are under purifying selection and tend to maintain their required functions. Meanwhile, the only gene observed to be under diversifying selection is *rpl23*. This gene has been reported to be deleted, duplicated, and accumulate mutations not only in legumes but also in other plant species, including cereal crops [29, 94–100].

Among the 64 codons directing protein synthesis, 61 encode standard amino acids, while 3 serve as translation stop signals. Most amino acids have multiple synonymous codons, except for tryptophan and methionine, typically encoded by one codon each [101]. The degeneracy of the genetic code allows the same amino acid to be encoded by different codons [102, 103]. However, codon usage varies among organisms, genes, and even the same gene from different species, resulting in codon usage bias [104, 105]. Codon usage bias leads to non-random appearance of synonymous codons with different frequencies [106, 107]. Codon bias impacts numerous cellular processes, such as mRNA stability, transcription, translation efficiency, and protein expression and cotranslation folding [108–110]. It influences chromatin structure and mRNA folding, thereby regulating transcription levels and translation efficiency by modulating the elongation rate of translation [108–111]. Codon bias analysis aids in revealing horizontal gene transfer and evolutionary relationships between closely related organisms [112, 113]. The RSCU value compares the observed frequency of a specific synonymous codon to the expected frequency (no codon usage bias). A value of 1.0 suggests no bias, with equal codon usage for that amino acid. Values above 1.0 indicate positive bias, while those below 1.0 indicate negative bias. RSCU values exceeding 1.6 or falling below 0.6 indicate overrepresented and underrepresented codons, respectively [38, 114].

Excluding the termination codon, we found 61 codons for 20 amino acids in the winged bean chloroplast genome, with RSCU values ranging from 0.47 to 4.38. Notably, CGC, encoding alanine, exhibited the lowest RSCU, while TTT, encoding phenylalanine, showed the highest RSCU. Phenylalanine, lysine, and asparagine were the most abundant amino acids encoded by the winged bean genome. Among the various synonymous codons for amino acids such as arginine, asparagine, aspartic acid, glutamic acid, isoleucine, leucine, lysine, serine, and tyrosine in the winged bean chloroplast genome, codons ending with either A or U were overrepresented. This bias towards codons ending with A or U may be attributed to the higher AT content of chloroplast genomes, resulting from mutation and natural selection processes [38, 115].

Understanding repeat sequences within genomes is critical for deciphering evolutionary patterns and genetic diversity [116, 117]. In the present study, we identified 59 repeats, primarily concentrated in intergenic spacers, consistent with patterns in other legumes studied recently [82, 118, 119]. A large proportion of repeats were also

found in the genes namely *ycf2, ndhA, ndhF, rps12, rpl22, pafI and psaA*, indicating potential functional implications. The prevalence of repeats in intergenic regions highlights their role in genomic rearrangements and evolutionary dynamics [120, 121]. Leveraging these conserved patterns as genetic markers can enhance phylogenetic and population studies in legumes, providing insights into chloroplast genome evolution and plant adaptation [120–122].

We identified 84 perfect SSRs, two compound SSRs, and 15 VNTRs in the chloroplast genome of winged bean. Their distribution in the LSC, SSC, and IR regions was generally similar to that observed in other legumes [99, 82, 118, 123, 124,125]. Typically, a significant proportion of SSRs in genic regions consist of trinucleotide repeats, which help mitigate the detrimental effects of frame-shift mutations [126, 127]. However, in our study, we found only mono- and dinucleotide repeats, predominantly concentrated in intergenic spacers and introns rather than exons. This may help counteract the detrimental effects of frame-shift mutations caused by mono- and dinucleotide repeats in genic regions [128–130]. The SSR markers identified in the chloroplast genome of the winged bean are of significant advantage in evolutionary and taxonomic research due to their maternal inheritance and lower mutation rates [126, 127].

In conclusion, the study of the chloroplast genome of *P. tetragonolobus* and its comparative analysis with other legumes has provided valuable insights into the evolutionary dynamics and phylogenetic relationships within the legume family. The findings highlight the Andrews, utility of chloroplast genome sequences in deep phylogenetic analysis and support the strength of the cladistic approach to biological classification based on the hypotheses of most recent common ancestry. The observed genomic homogeneity among the leguminous species under investigation and the uniformity in GC content across different regions underscores the conserved nature of chloroplast genomes within this plant family. These results contribute to our understanding of legume systematics and emphasise the importance of combining expertise in genetics and bioinformatics to capitalise on new approaches for studying plant evolutionary biology.

Declarations

Acknowledgment

We acknowledge the funding and logistic support provided by Director, ICAR-IIAB, Ranchi.

Authors Contribution

KUT – Sequencing, data analysis, NKS – Data analysis, AG – Data analysis, HS - Data analysis and submission of sequences to NCBI, AP- Data analysis, SK- Generated the plant material, VPB-Coordinated the project, SR-Coordinated the project and manuscript editing, APT- Coordinated the project, BKS- Planning the experiment, sequencing and writing the manuscript.

Conflict of interest

The authors declare no conflict of interest.

Data availability statement

Sequence data that support the finding of this study have been deposited in NCBI BankIt with primary accession code Psophocarpus PP894786.1.

References

1. Bennett, M. D. & Smith, J. B. Nuclear dna amounts in angiosperms. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.***274**(933), 227–274 (1976).
2. Harder, D. K. & Smartt, J. Further evidence on the origin of the cultivated winged bean, *Psophocarpus tetragonolobus* (L.) DC. (Fabaceae): Chromosome numbers and the presence of a host-specific fungus. *Econ. Bot.***46**, 187–191 (1992).
3. Khan, T. N. Papua New Guinea: A centre of genetic diversity in winged bean (*Psophocarpus tetragonologus* (L.) DC.). *Euphytica***25**, 693–705 (1976).
4. Sriwichai, S. *et al.* Genetic diversity of domestic (Thai) and imported winged bean [*Psophocarpus tetragonolobus* (L.) DC.] cultivars assessed by morphological traits and microsatellite markers. *Ann. Agric. Sci.***67**(1), 34-41 (2022).
5. Khan, T. N., Bohn, J. C. & Stevenson, R. A. Winged bean: Cultivation in Papua New Guinea. *World Crops***29**(5), 208–214 (1977).
6. Banerjee, A., Bagchi, D. & Si, L. K. Studies on the potential of winged bean as a multipurpose legume cover crop in tropical regions. *Exp. Agric.***20**(4), 297-301 (1984).
7. Anugroho, F., Kitou, M., Kinjo, K. & Kobashigawa, N. Growth and nutrient accumulation of winged bean and velvet bean as cover crops in a subtropical region. *Plant Prod. Sci.***13**, 360–366 (2010).
8. Masefeld, G. B. Root nodulation and agricultural potential of the leguminous genus Psophocarpus. *Trop. Agric.***38**, 225–229 (1961).
9. Haq, N. Germplasm resources, breeding and genetics of the winged bean. *Zeitschrift fur Pflanzenzuchtung - Journal of plant breeding***88**(1), 1-12 (1982).
10. Makeri, M. U. *et al.* Comparative physiochemical, functional and structural characteristics of winged bean [*Psophocarpus tetragonolobus* DC] and soybean [*Glycine max*] protein isolates. *J. Food Meas. Charact.* **11**, 835-846 (2017).
11. Mohanty, C. S. *et al.* Physicochemical analysis of *Psophocarpus tetragonolobus* (L.) DC seeds with fatty acids and total lipids compositions. *J. Food Sci. Technol.***52**(6), 3660– 3670 (2015).
12. Makeri, M. U. *et al.* Comparative analysis of the physiochemical, thermal, and oxidative properties of winged bean and soybean oils. *Int. J. Food Prop.***19**(12), 2769–2787 (2016).
13. Oso, A. A. & Ashafa, A. O. Nutritional composition of grain and seed proteins in *Grain and seed proteins functionality* (ed. Jimenez-Lopez, J.C.) (IntechOpen, 2021).
14. Lepcha, P., Egan, A. N., Doyle, J. J. & Sathyanarayana, N. A review on current status and future prospects of winged bean (*Psophocarpus tetragonolobus*) in tropical agriculture. *Plant Foods Hum. Nutr.* **72**(3), 225-235 (2017).
15. Verdcourt, B. & Halliday, P. A revision of *Psophocarpus* (Leguminosae-Papilionoideae -Phaseoleae). *Kew Bull.* **33**, 191; 10.2307/41095 75 (1978).
16. Schimper, A. Uber die entwicklung der chlorophyllkorner und farbkorper. *Bot. Zeitung***30**, 105–114, 121–131, 137–146, 153–162 (1883).

17. Martin, W. *et al.* Evolutionary analysis of arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc. Natl. Acad. Sci. USA*. **99**, 12246–12251 (2002).
18. Palmer, J. D. Comparative organisation of chloroplast genomes. *Annu. Rev. Genet*. **19**, 325–354 (1985).
19. Wang, R. J. *et al.* Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. *BMC Evol. Biol.***8**, 36; 10.1186/1471-2148-8-36 (2008).
20. Maier, R. M., Neckermann, K., Igloi, G. L. & Kössel, H. Complete sequence of the maize chloroplast genome: Gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. *J. Mol. Biol*. **251**, 614–628 (1995).
21. Jansen, R. K. & Ruhlman, T. A. Plastid genomes of seed plants in *Genomics of chloroplasts and mitochondria: Advances in photosynthesis and respiration* (eds. Bock, R. & Knoop V.) 35, 103-126 (Springer Publishers, The Netherlands, 2012).
22. Jansen, R. K. *et al.* 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.***48**, 1204–1217 (2008).
23. Wicke, S. & Schneeweiss, G. M. Next-generation organellar genomics: Potentials and pitfalls of high-throughput technologies for molecular evolutionary studies and plant systematics in *Next-generation sequencing in plant systematics* (eds. Hörandl, E. & Appelhans, M.) 1-42 (International Association for Plant Taxonomy (IAPT), Bratislava, Slovakia, 2015).
24. Palmer, J. D. & Thompson, W. F. Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. *Cell***29**, 537–550 (1982).
25. Lavin, M., Doyle, J. J. & Palmer, J. D. Evolutionary significance of the loss of the chloroplast-DNA inverted repeat in the Leguminosae subfamily Papilionoideae. *Evolution***44**(2), 390-402 (1990).
26. Bruneau, A. *et al.* Legume phylogeny and classification in the 21st century: Progress, prospects and lessons for other species-rich clades. *Taxon* **62**, 217-248 (2013).
27. Doyle, J. J., Doyle, J. L. & Palmer, J. D. Multiple independent losses of two genes and one intron from legume chloroplast genomes. *Syst. Bot*. **20**, 272–294 (1995).
28. Gantt, J. S., Baldauf, S. L., Calie, P. J., Weeden, N. F. & Palmer, J. D. Transfer of *rpl22* to the nucleus greatly preceded its loss from the chloroplast and involved the gain of an intron. *EMBO J*. **10**, 3073–3078 (1991).
29. Magee, A. M. *et al.* Localised hypermutation and associated gene losses in legume chloroplast genomes. *Genome Res.* **20**, 1700–1710 (2010).
30. Moore, M. J., Soltis, P. S., Bell, C. D., Burleigh, J. G. & Soltis, D. E. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proc. Natl. Acad. Sci. USA*. **107**, 4623–4628 (2010).
31. Zhang, D., Li, K., Gao, J., Liu, Y. & Gao, L.-Z. The complete plastid genome sequence of the wild rice *Zizania latifolia* and comparative chloroplast genomics of the rice tribe Oryzeae, Poaceae. *Front. Ecol. Evol*. **4**; 10.3389/fevo.2016.00088 (2016).
32. Osuna-Mascaró, C., Rafael, R. D. C. & Perfectti, F. Comparative assessment shows the reliability of chloroplast genome assembly using RNA-seq. *Sci. Rep.***8**, 17404; 10.1038/s41598-018-35654-3 (2018).
33. Lewis, G. P. Legumes of the world, Royal Botanic Gardens, Kew. https://books.google.co.in/books?id=iC9FAQAAIAAJ (2005).

34. Käss, E. & Wink, M. Phylogenetic relationships in the Papilionoideae (family Leguminosae) based on nucleotide sequences of cpDNA (*rbcL*) and ncDNA (ITS 1 and 2). *Mol. Phylogenet. Evol.***8**, 65–88 (1997).
35. Brouat, C., Gielly, L. & McKey, D. Phylogenetic relationships in the genus *Leonardoxa* (Leguminosae: Caesalpinioideae) inferred from chloroplast *trnL* intron and *trnL-trnF* intergenic spacer sequences. *Am. J. Bot.* **88**, 143–149 (2001).
36. Manzanilla, V. & Bruneau, A. Phylogeny reconstruction in the Caesalpinieae grade (Leguminosae) based on duplicated copies of the sucrose synthase gene and plastid markers. *Mol. Phylogenet. Evol.***65**, 149–162 (2012).
37. Zhou, T. *et al*. Comparative transcriptome and chloroplast genome analyses of two related Dipteronia species. *Front. Plant Sci.***7**, 1512; 10.3389/fpls.2016.01512 (2016).
38. Parvathy, S. T., Udayasuriyan, V. & Bhadana, V. Codon usage bias. *Mol. Biol. Rep.***49**(1), 539–565 (2022).
39. Somaratne, Y., Guan, D. L., Wang, W. Q., Zhao, L. & Xu, S. Q. The complete chloroplast genomes of two *Lespedeza* species: Insights into codon usage bias, RNA editing sites, and phylogenetic relationships in desmodieae (Fabaceae: Papilionoideae). *Plants***9**(1), 51; 10.3390/plants9010051 (2019).
40. Wang, M. *et al.* Comparative analysis of *Asteraceae* chloroplast genomes: Structural organisation, RNA editing and evolution. *Plant Mol. Biol. Rep.* **33**, 1526–1538 (2015).
41. DeCosa, B., Moar, W., Lee, S. B., Miller, M. & Daniell, H. Overexpression of the *Bt cry2Aa2* operon in chloroplasts leads to formation of insecticidal crystals. *Nat. Biotechnol.* **19**, 71–74 (2001).
42. Ruiz, O. N., Hussein, H., Terry, N. & Daniell, H. Phytoremediation of organomercurial compounds via chloroplast genetic engineering. *Plt. Phys.* **132**, 1344–1352 (2003).
43. Hagemann, R. The sexual inheritance of plant organelles in *Molecular biology and biotechnology of plant organelles* (eds. Daniell, H. & Chase, C. D.) 93-113 (Springer Publishers, Dordrecht, The Netherlands, 2004).
44. Dhingra, A., Portis, A. R. & Daniell, H. Enhanced translation of a chloroplast expressed *rbcS* gene restores SSU levels and photosynthesis in nuclear antisense *RbcS* plants. *Proc. Natl. Acad. Sci. USA.* **101**, 6315–6320 (2004).
45. Daniell, H., Khan, M. & Allison, L. Milestones in chloroplast genetic engineering: An environmentally friendly era in biotechnology. *Trends Plant Sci.***7**, 84–91 (2002).
46. Leelavathi, S., Gupta, N., Maiti, S., Ghosh, A. & Reddy, V. S. Overproduction of an alkali- and thermo-stable xylanase in tobacco chloroplasts and efficient recovery of the enzyme. *Mol. Breed.***11**, 59–67 (2003).
47. Daniell, H., Cohill, P. R., Kumar, S. & Dufourmantel, N. Chloroplast genetic engineering in *Molecular biology and biotechnology of plant organelles* (eds. Daniell, H. & Chase, C. D.) 443–490 (Springer Publishers, The Netherlands, 2004).
48. Daniell, H., Kumar, S. & Duformantel, N. Breakthrough in chloroplast genetic engineering of agronomically important crops. *Trends Biotechnol.***23**, 238–245 (2005).
49. Dhingra, A. & Folta, K. M. ASAP: Amplification, sequencing & annotation of plastomes. *BMC Genom.***6**, 176; 10.1186/1471-2164-6-176 (2005).
50. Cronn, R. *et al.* Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Res.* **36**(19), e122; 10.1093/nar/gkn502 (2008).
51. Dong, W., Xu, C., Cheng, T., Lin, K. & Zhou, S. Sequencing angiosperm plastid genomes made easy: A complete set of universal primers and a case study on the phylogeny of saxifragales, *Genome Biol. Evol.* **5**, 989–997 (2013).

52. Moore, M. J. *et al.* Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biol.* 6, 17; 10.1186/1471-2229-6-17 (2006).

53. Kaila, T. *et al.* Chloroplast genome equence of pigeonpea (*Cajanus cajan* (L.) Millspaugh) and *Cajanus scarabaeoides* (L.) Thouars: Genome organisation and comparison with other legumes. *Front. Plant Sci.* **7**, 1847; 10.3389/fpls.2016.01847 (2016).

54. Jordan, W. C., Courtney, M. W. & Neigel, J. E. Low levels of intraspecific genetic variation at a rapidly evolving chloroplast DNA locus in North American duckweeds (Lemnaceae). *Am. J. Bot.* **83**, 430–439 (1996).

55. Uthaipaisanwong, P. *et al.* Characterisation of the chloroplast genome sequence of oil palm (*Elaeis guineensis* Jacq). *Gene***500**, 172–180 (2012).

56. Yang, H. *et al.* Genetic divergence between *Camellia sinensis* and its wild relatives revealed *via* genome-wide SNPs from RAD sequencing. *PLoS One***11**, e0151424; 10.1371/journal.pone.0151424 (2016).

57. Murray, M. G. & Thompson, W. F. Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* **8**(19), 4321-4325 (1980).

58. Chen, S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta***2**(2), e107; 10.1002/imt2.107 (2023).

59. Jin, J. J. *et al.* GetOrganelle: A fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. *Genome Biol.* **21**, 241; 10.1186/s13059-020-02154-5 (2020a).

60. Tillich, M. *et al.* GeSeq - Versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.***45**(W1),W6–W11 (2017).

61. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**(4), 772–780 (2013).

62. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics***35**(21), 4453–4455 (2019).

63. Rabiee, M., Sayyari, E. & Mirarab, S. Multi-allele species reconstruction using ASTRAL. *Mol. Phylogenetics Evol.* **130**, 286–296 (2019).

64. Guy, L., Kultima, J. R. & Andersson, S. G. E. genoPlotR: Comparative gene and genome visualisation in R. *Bioinformatics***26**(18), 2334–2335 (2010).

65. Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M. & Dubchak, I. VISTA: Computational tools for comparative genomics. *Nucleic Acids Res.* **32**(Web Server issue), W273-9; 10.1093/nar/gkh458 (2004).

66. Katoh, K., Standley, D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30(4):772-80. doi: 10.1093/molbev/mst010 (2013)

67. Zhang, Z. KaKs_Calculator 3.0: Calculating selective pressure on coding and non-coding sequences. *Genom. Proteom. & Bioinf.***20**(3), 536–540 (2022).

68. Wickham, H. Ggplot2 : Elegant graphics for data analysis. Springer-Verlag New York. https://tidyverse.github.io/ggplot2-docs/authors.html (2016).

69. Kurtz, S. *et al.* REPuter: The manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.***29**, 4633–4642 (2001).

70. Du, L., Zhang, C., Liu, Q., Zhang, X. & Yue, B. Krait: an ultrafast tool for genome-wide survey of microsatellites and primer design. *Bioinformatics***34**(4), 681-683 (2018).

71. Magallón, S. & Sanderson, M. J. Absolute diversification rates in angiosperm clades. *Evolution***55**, 1762–1780 (2001).
72. Klu, G. Y. P. & Kumaga, F. K. Testing of induced mutants of winged bean (*Psophocarpus tetragonolobus* (L.) DC.) for nodulation and phenotypic performance. *Ghana J. Sci.***39**, 55–62 (1999).
73. Bock, R. & Volker K. Genomics of chloroplasts and mitochondria (Springer Science & Business Media, 2012).
74. Wicke, S., Schneeweiss, G. M., dePamphilis, C. W., Müller, K. F. & Dietmar, Q. The evolution of the plastid chromosome in land plants: Gene content, gene order, gene function. *Plant Mol. Biol.* **76**(3-5), 273–297 (2011).
75. Bruneau, A., Mercure, M., Lewis, G. P. & Herendeen, P. S. Phylogenetic patterns and diversification in the caesalpinioid legumes. *Botany***86**, 697–718 (2008).
76. Lavin, M. *et al.* The Dalbergioid legumes (Fabaceae): Delimitation of a pantropical monophyletic clade. *Am. J. Bot.* **88** (3), 503–533 (2001).
77. Oyebanji, O., Rong, Z., Chen, S.-Y. & Yi, T.-S. New insights into the plastome evolution of the Millettioid/Phaseoloid clade (Papilionoideae, Leguminosae). *Front. Plant Sci.* **11**; 10.3389/fpls.2020.00151 (2020).
78. Schrire, B. D., Matt, L., Nigel, P. B. & Félix, F. Phylogeny of the tribe Indigofereae (Leguminosae-Papilionoideae): Geographically structured more in succulent-rich and temperate settings than in grass-rich environments. *Am. J. Bot.* **96**(4), 816–852 (2009).
79. Wojciechowski, M. F., Matt L. & Sanderson, M. J. A phylogeny of legumes (Leguminosae) based on analysis of the plastid *matK g*ene resolves many well-supported subclades within the family. *Am. J. Bot.* **91**(11), 1846–1862 (2004).
80. Choi, I.-S. *et al*. Highly resolved Papilionoid legume phylogeny based on plastid phylogenomics. *Front. Plant Sci.* **13**, 823190; 10.3389/fpls.2022.823190 (2022).
81. Brottier, L. *et al.* A phylogenetic framework of the legume genus *Aeschynomene* for comparative genetic analysis of the nod-dependent and nod-independent symbioses. *BMC Plant Biol.* **18**, 333; 10.1186/s12870-018-1567-z (2018).
82. Mehmetoğlu, E., Kaymaz, Y., Ateş, D., Kahraman, A. & Tanyolaç, M. B. The complete chloroplast genome of *Cicer reticulatu*m and comparative analysis against relative *Cicer* species. *Sci. Rep.***13**, 17871; 10.1038/s41598-023-44599-1 (2023).
83. Chen, R. Functional genomics and genetic control of compound leaf development in *Medicago Truncatula*: An overview. *Methods Mol. Biol.***1822**, 197–203 (2018).
84. Duan, L. *et al.* Phylogenomic framework of the IRLC legumes (Leguminosae Subfamily Papilionoideae) and intercontinental biogeography of tribe Wisterieae. *Mol. Phylogenetics Evol.* **163**, 107235; 10.1016/j.ympev.2021.107235 (2021).
85. Bowers, J. E., Haibao, T., John, M. B. & Andrew, H. P. GC content of plant genes is linked to past gene duplications. *PloS One* **17** (1), e0261748; 10.1371/journal.pone.0261748 (2022).
86. Muyle, A., Laurana S.-G., Adrienne, R., Juan, E. & Sylvain, G. GC-biased gene conversion and selection affect GC content in the *Oryza* genus (rice). *Mol. Biol. Evol.* **28**(9), 2695–2706 (2011).
87. Trávníček, P. *et al.* Diversity in genome size and GC content shows adaptive potential in orchids and is closely linked to partial endoreplication, plant life-history traits and climatic conditions. *New Phytol.* **224**(4), 1642–56 (2019).

88. Ogihara, Y., Terachit, T. & Sasakuma, T. Intramolecular recombination of chloroplast genome mediated by short direct-repeat sequences in wheat species (hot spot/DNA length mutation/genetic diversity of plastome/Tritcum/Aegiops). *Proc. Natl. Acad. Sci. USA.***85**, 8573–8577 (1988).
89. Jin, D. M., Jin, J. J. & Yi, T. S. Plastome structural conservation and evolution in the Clusioid clade of Malpighiales. *Sci. Rep.***10**, 9091; 10.1038/s41598-020-66024-7 (2020b).
90. Zhou, S., Ma, K., Mower, J. P., Liu, Y. & Zhou, R. Leaf variegation caused by plastome structural variation: An example from *Dianella tasmanica*. *Hortic. Res.***11**(3), uhae009 10.1093/hr/uhae009 (2024).
91. Hurst, L. D. The Ka/Ks ratio: Diagnosing the form of sequence evolution. *Trends Genet.* **18**(9), 486–487 (2002).
92. Yang, Z. & Bielawski, J. R. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.***15**(12), 496–503 (2000).
93. Howe, C. J. *et al.* Evolution of the chloroplast genome. *Philos. Trans. R. Soc. B: Biol. Sci.* **358**(1429), 99–107 (2003).
94. Claude, S. J., Park, S., & Park, S. J. Gene loss, genome rearrangement, and accelerated substitution rates in plastid genome of *Hypericum ascyron* (Hypericaceae). *BMC Plant Biol.* **22**(1), 1–12; 10.1186/S12870-022-03515-X/FIGURES/6 (2022).
95. Daniell, H., Lin, C. S., Yu, M. & Chang, W. J. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biol.* **17** (1), 1–29 (2016).
96. Lencina, F. *et al.* The *rpl23* gene and pseudogene are hotspots of illegitimate recombination in barley chloroplast mutator seedlings. *Sci. Rep.***9**, 9960; 10.1038/s41598-019-46321-6 (2019).
97. Mohanta, T. K. *et al.* Gene loss and evolution of the plastome. *Genes* **11**(10), 1133; 10.3390/GENES11101133 (2020).
98. Souza, U. J. B., Nunes, R., Targueta, C. P., Diniz-Filho, J. A. F., Telles, M. P. C. The complete chloroplast genome of *Stryphnodendron adstringens* (Leguminosae - Caesalpinioideae): Comparative analysis with related Mimosoid species. *Sci. Rep.* **9**, 14206; 10.1038/s41598-019-50620-3 (2019).
99. Tian, S. *et al.* Chloroplast genome sequence of Chongming lima bean (*Phaseolus Lunatus* L.) and comparative analyses with other legume chloroplast genomes. *BMC Genom.* **22**(1), 194; 10.1186/s12864-021-07467-8 (2021).
100. Zhou, S. M. *et al.* Phylogenomics and plastome evolution of Indigofera (Fabaceae). *Front. Plant Sci.***14**, 1186598; 10.3389/fpls.2023.1186598 (2023).
101. Quax, T. E., Claassens, N. J., Soll, D. & van der Oost, J. Codon bias as a means to fine-tune gene expression. *Mol. Cell***59**, 149–161 (2015).
102. Yao, H., Chen, M. & Tang, Z. Analysis of synonymous codon usage bias in Flaviviridae virus. *Biomed. Res. Int.* 5857285; 10.1155/2019/5857285 (2019).
103. Bailey, S. F., Alonso Morales, L. A. & Kassen, R. Effects of synonymous mutations beyond codon bias: The evidence for adaptive synonymous substitutions from microbial evolution experiments. *Genome Biol. Evol.***13**, evab141; 10.1093/gbe/evab141 (2021).
104. Behura, S. K. & Severson, D. W. Comparative analysis of codon usage bias and codon context patterns between dipteran and hymenopteran sequenced genomes. *PLoS One***7**, e43111; 10.1371/journal.pone.0043111 (2012).

105. Barbhuiya, P. A., Uddin, A. & Chakraborty, S. Analysis of compositional properties and codon usage bias of mitochondrial *CYB* gene in anura, urodela and gymnophiona. *Gene***751**, 144762; 10.1016/j.gene.2020.144762 (2020).

106. Hershberg, R. & Petrov, D. A. Selection on codon bias. *Annu. Rev. Genet.***42**, 287–299 (2008).

107. Choudhury, M. N., Uddin, A. & Chakraborty, S. Nucleotide composition and codon usage bias of *SRY* gene. *Andrologia***50**, e12787; 10.1111/and.12787 (2018).

108. Chu, D. & Wei, L. Direct *in vivo* observation of the effect of codon usage bias on gene expression in *Arabidopsis* hybrids. *J. Plant Physiol.***265**, 153490; 10.1016/j.jplph.2021.153490 (2021).

109. Franzo, G., Tucciarone, C. M., Legnardi, M. & Cecchinato, M. Effect of genome composition and codon bias on infectious bronchitis virus evolution and adaptation to target tissues. *BMC Genom.* **22**, 244; 10.1186/s12864-021-07559-5 (2021).

110. Hia, F. & Takeuchi, O. The effects of codon bias and optimality on mRNA and protein regulation. *Cell. Mol. Life Sci.***78**, 1909–1928 (2021).

111. Uddin, A., Mazumder, T. H., Choudhury, M. N. & Chakraborty, S. Codon bias and gene expression of mitochondrial *ND2* gene in chordates. *Bioinformation***11**, 407–412 (2015).

112. LaBella, A. L., Opulente, D. A., Steenwyk, J. L., Hittinger, C. T. & Rokas, A. Variation and selection on codon usage bias across an entire subphylum. *PLoS Genet.* 15, e1008304; 10.1371/journal.pgen.1008304 (2019).

113. Kokate, P. P., Techtmann, S. M. & Werner, T. Codon usage bias and dinucleotide preference in 29 drosophila species. G3 **11**, jkab191; 10.1093/g3journal/jkab191 (2021).

114. Iriarte, A., Guillermo, L. & Héctor, M. Codon usage bias: An endless tale. *J. Mol. Evol.* **89** (9-10), 589–93 (2021).

115. Hockenberry, A. J., Sirer, M. I., Amaral, L. A. N. & Michael C. J. Quantifying position-dependent codon usage bias. *Mol. Biol. Evol.* **31**(7), 1880–1893 (2014).

116. Qin, Z. *et al.* Evolution analysis of simple sequence repeats in plant genome. *PloS One* **10**(12), e0144108; 10.1371/journal.pone.0144108.10 (2015).

117. Wheeler, G. L., Hanna, E. D., Alenda, B., Lavanya, C. & Lisa, E. W. A review of the prevalence, utility, and caveats of using chloroplast simple sequence repeats for studies of plant biology. *Appl. Plant Sci.* **2**(12), apps.1400059; 10.3732/apps.1400059 (2014).

118. Wu, L. *et al.* Comparative genomics and phylogenomics of the genus *Glycyrrhiza* (Fabaceae) based on chloroplast genomes. *Front. Pharmacol.* **15**, 1371390; 10.3389/fphar.2024.1371390 (2024).

119. Zhang, T. *et al.* Comparative analysis of chloroplast pan-genomes and transcriptomics reveals cold adaptation in *Medicago sativa*. *Int. J. Mol. Sci.***25**(3), 1776; 10.3390/ijms25031776 (2024).

120. Štorchová, H. & Krüger, M. The overview of methods for assembling complex mitochondrial genomes in land plants. *J. Exp. Bot.* erae034; 10.1093/jxb/erae034 (2024).

121. Cao, J. *et al.* Extreme reconfiguration of plastid genomes in *Papaveraceae*: Rearrangements, gene loss, pseudogenization, IR expansion, and repeats. *Int. J. Mol. Sci.***25**(4) 2278; 10.3390/ijms25042278 (2024).

122. Wang, J. *et al.* Plant organellar genomes: Much done, much more to do. *Trends Plant Sci.* https://doi.org/10.1016/j.tplants.2023.12.014 (2024)

123. Park, J.-S., Youngsung, K., Iseon, K., Ho-Rim, H. & Changkyun, K. Complete chloroplast genome of *Albizia kalkora* (Fabaceae) in Korea and its taxonomic position. *Journal of Asia-Pacific Biodiversity* **17**(2); 10.1016/j.japb.2023.12.009 (2024).

124. Yue, Y., Jiawen, L., Xuegang, S., Zhen L. & Bingjun, J. Polymorphism analysis of the chloroplast and mitochondrial genomes in soybean. *BMC Plant Biol.***23**, 15; 10.1186/s12870-022-04028-3 (2023).

125. Yang, S., Gun, L. & Hao, L. Molecular characterisations of genes in chloroplast genomes of the genus *Arachis L.* (Fabaceae) based on the codon usage divergence. *PloS One* **18**(3), e0281843; 10.1371/journal.pone.0281843 (2023).

126. Xin, D. *et al.* Identification and characterization of SSRs from soybean (*Glycine max*) ESTs. *Mol. Biol. Rep.***39**, 9047-9057 (2012).

127. Thumilan, M. *et al.* Development and characterization of genic SSR markers from Indian Mulberry transcriptome and their transferability to related species of Moraceae. *PLoS One***11** (9), e0162909; 10.1371/journal.pone.0162909 (2016).

128. Weisschuh, N., Elena, B.-A. & Bernd W. Splicing mutations in inherited retinal diseases. *Prog. Retin Eye Res.* **80**, 100874; 10.1016/j.preteyeres.2020.100874 (2021).

129. Zhu, L. *et al.* Short tandem repeats in plants: Genomic distribution and function prediction. *Electron. J. Biotechnol.* **50**, 37–44 (2021).

130. Xu, Y., Miaomiao, X., Lixiao, S., Jiyong, Y., Wenjiang, L. & Aisong, Z. Genome-wide analysis of simple sequence repeats in cabbage (*Brassica Oleracea* L.). *Front. Plant Sci.* **12**, 726084; 10.3389/fpls.2021.726084 (2021).

Figures

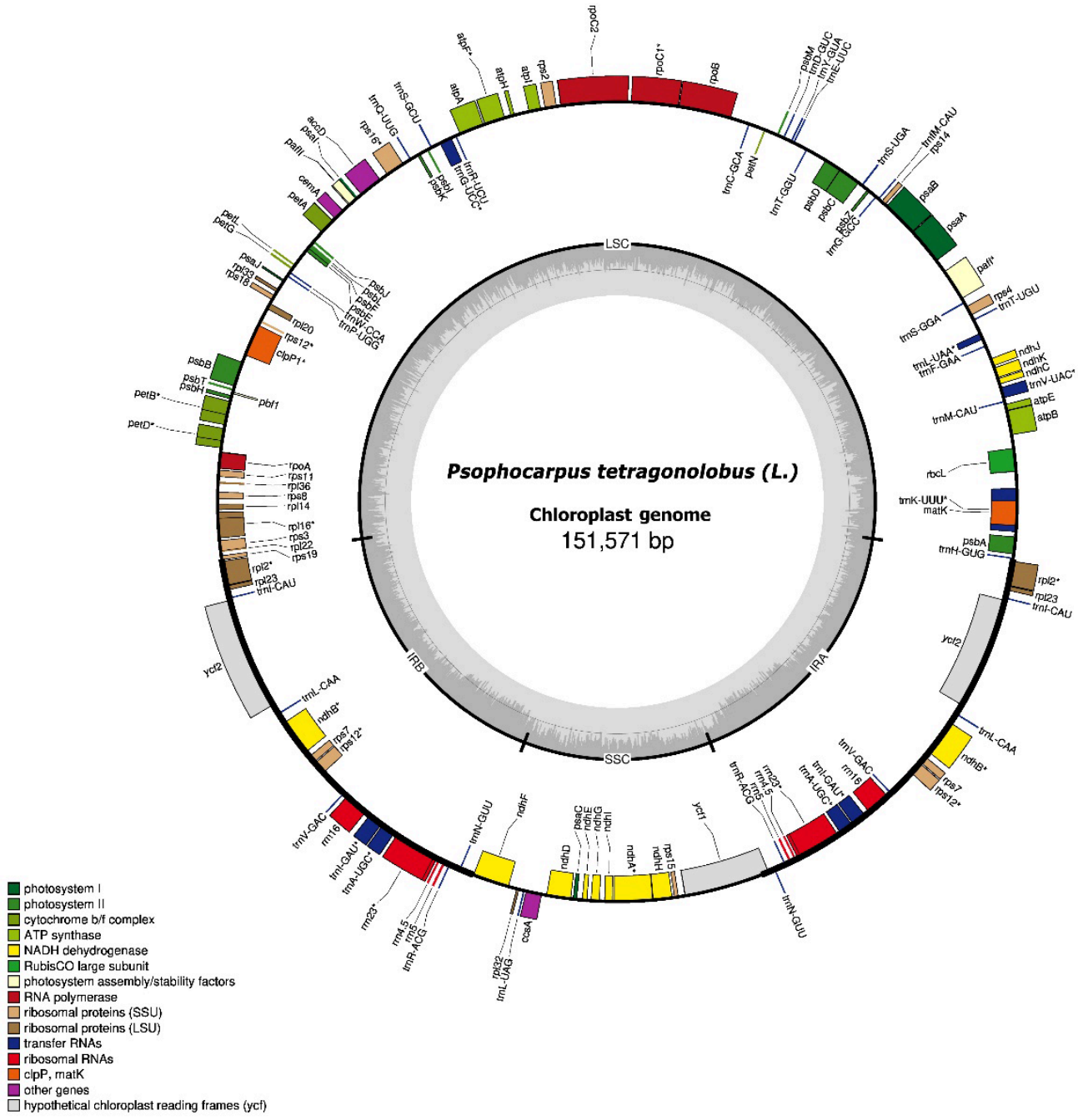


Figure 1

Chloroplast genome map of *Psophocarpus tetragonolobus*. Genes showed on the inner side are transcribed anti-clockwise while genes presented on the outer side are transcribed clockwise. Genes are color coded according to their functions as mentioned within the figure. The genome is divided into Inverted repeat B (IRB), Inverted repeat A (IRA), Small Single Copy (SSC) and Large Single Copy (LSC) region

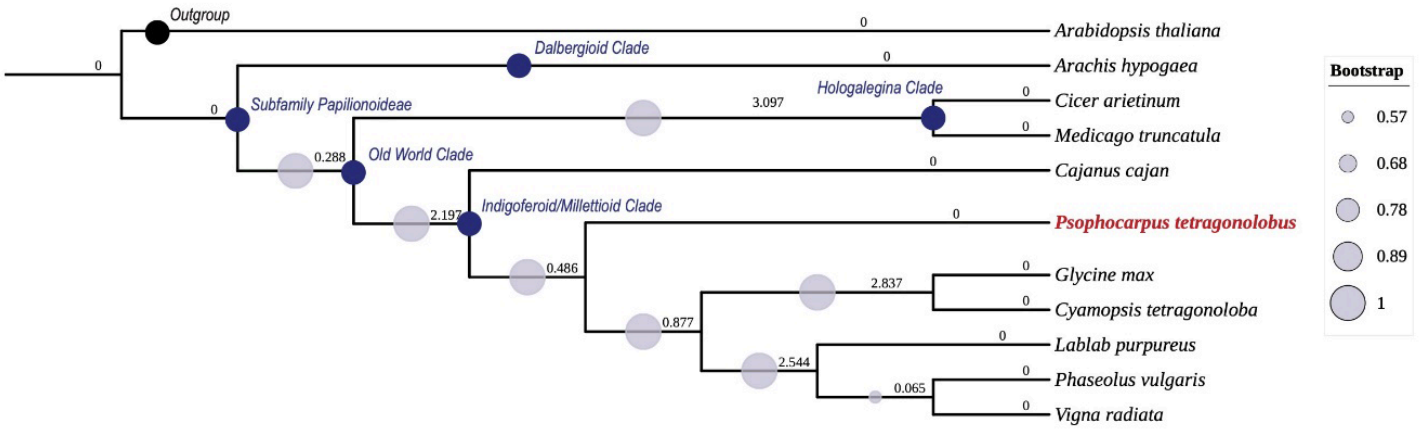


Figure 2

Gene-based phylogeny construction. Chloroplast genes (n=70) based on maximum likelihood Phylogenetic tree of 11 species. Respective node lengths are written above the branches, while bootstrap values are represented with weighted circular bubbles. The respective clade names are mentioned at the node of diversion, highlighted with dark blue circles.

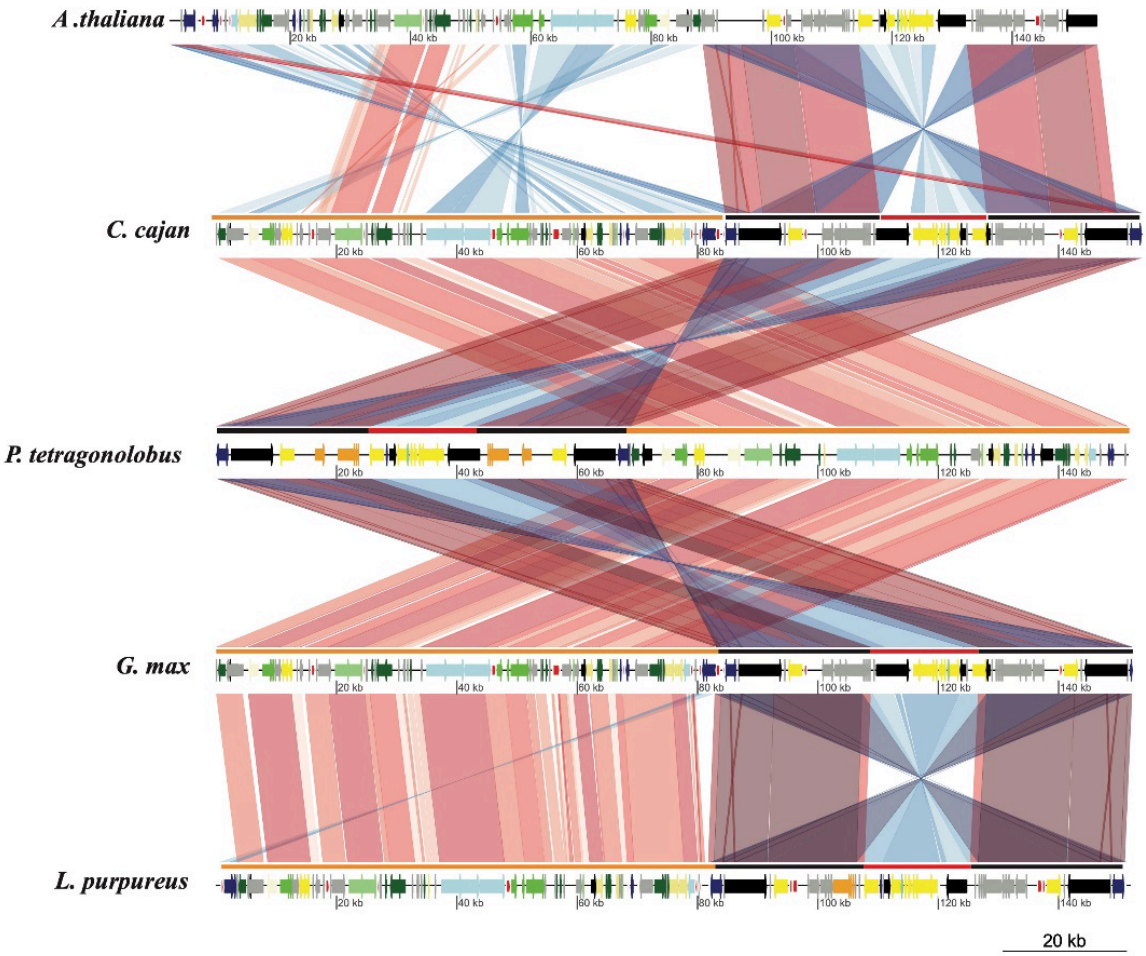
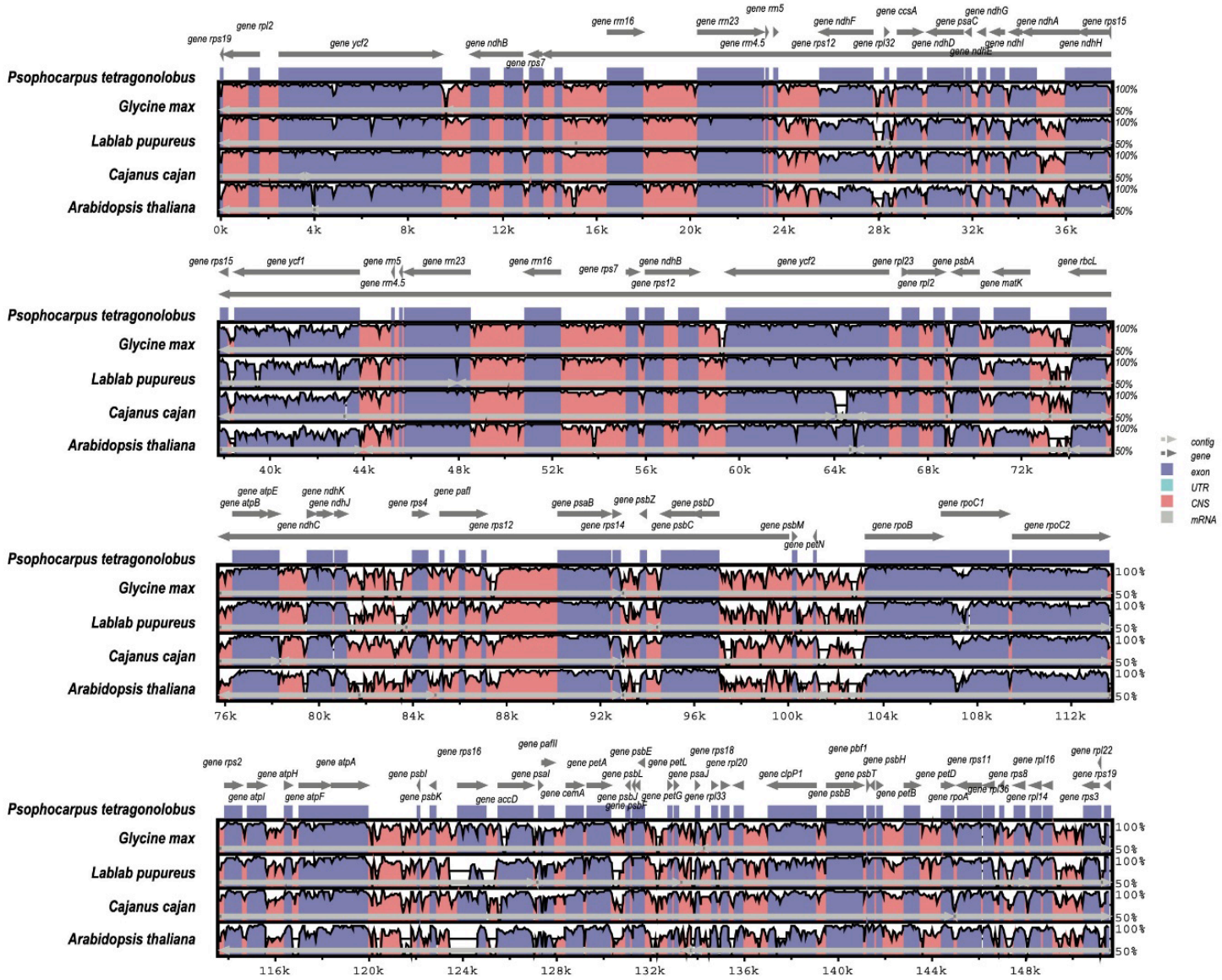


Figure 3



Synteny analysis between five species including *P.tetragonolobus*. Chloroplast synteny block analysis performed using BLASTn with the default parameters and visualised using genoplotR (Guy et al., 2010). Only syntenic regions with ≥100 bp are shown. Traces connecting cDNA genomes represent synteny blocks, while the gene arrowhead represents the forward or reverse direction of the gene. The genes share the same gene colour code as in genome figure 1. The red gradient segments represent the percentage of sequence identity from BLASTN alignments; thus, darker colors indicate higher identity.

Figure 4

Comparison of *Arabidopsis thalina* and three other legume species using whole chloroplast genome alignment. *P. tetragonolobus* was used as the reference for the alignment. The arrows indicate the length and direction of the gene. The color legends have been included within the figure.

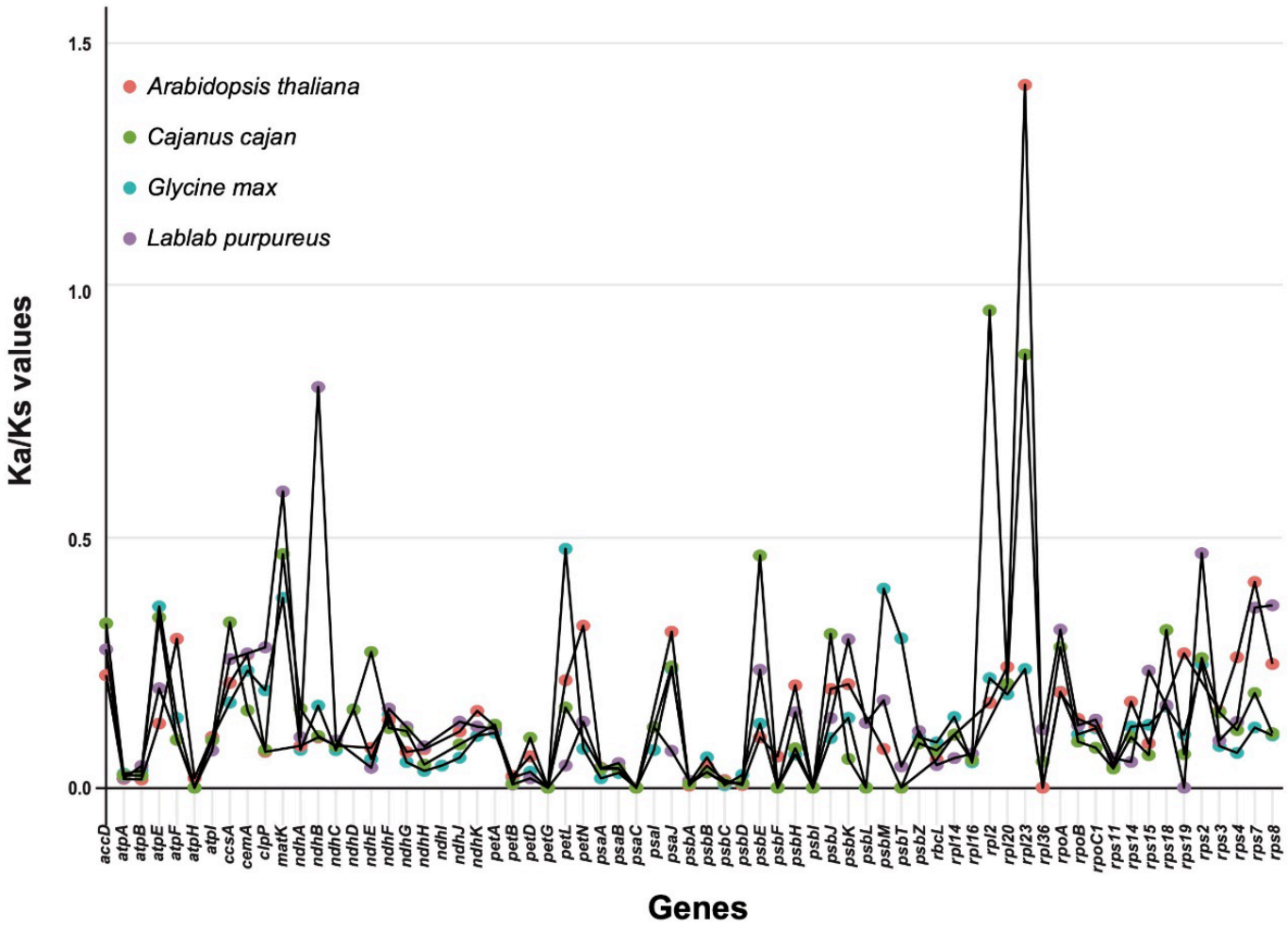


Figure 5

Distribution of non-synonymous and synonymous substitution within the gene coding sequences of winged bean when compared to three other legume species. *A.thaliana* was taken as an outgroup. Species are color coded and described within the figure.

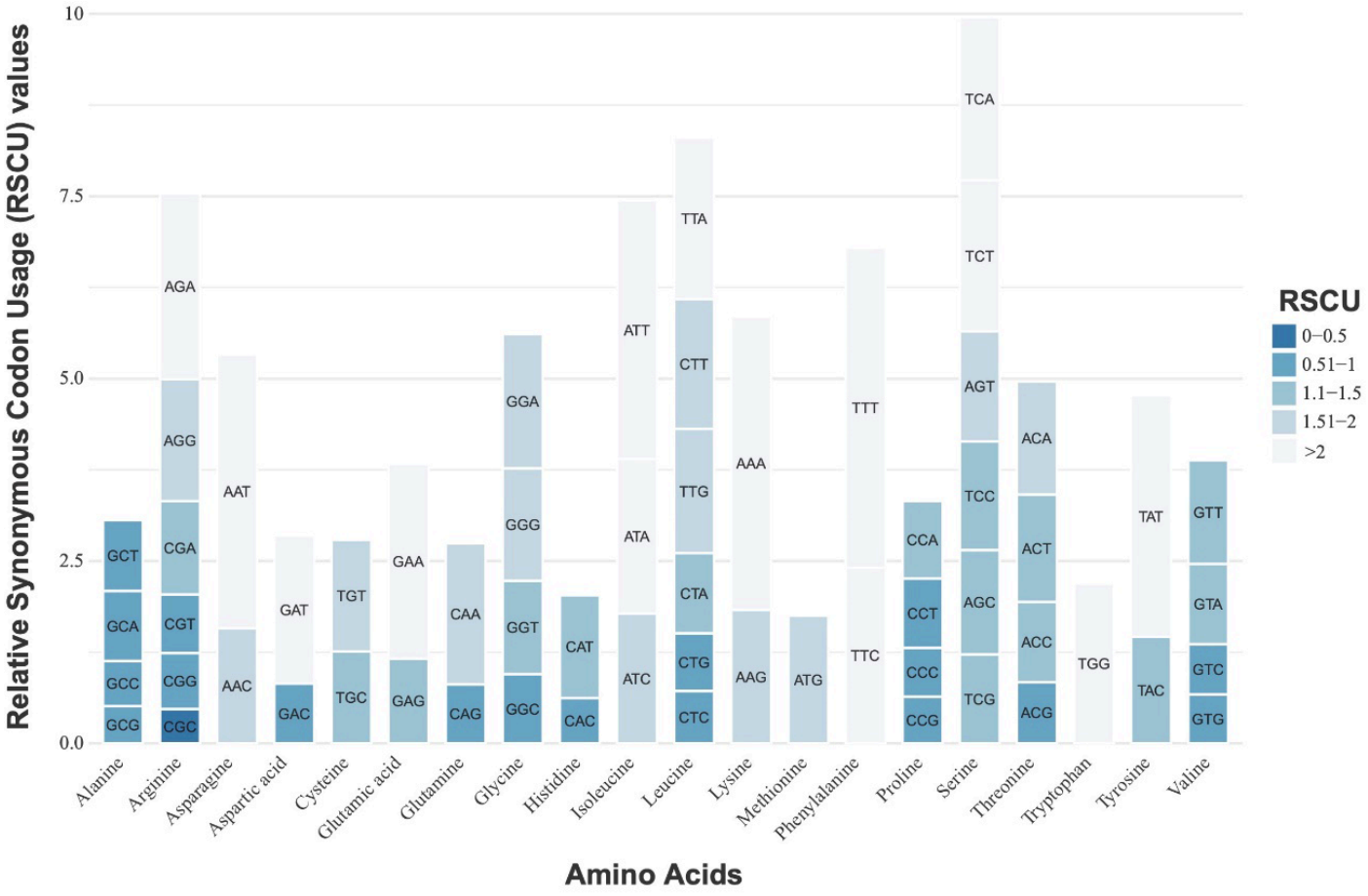


Figure 6

Relative synonymous codon usage (RSCU) for codons for coding their respective amino acids

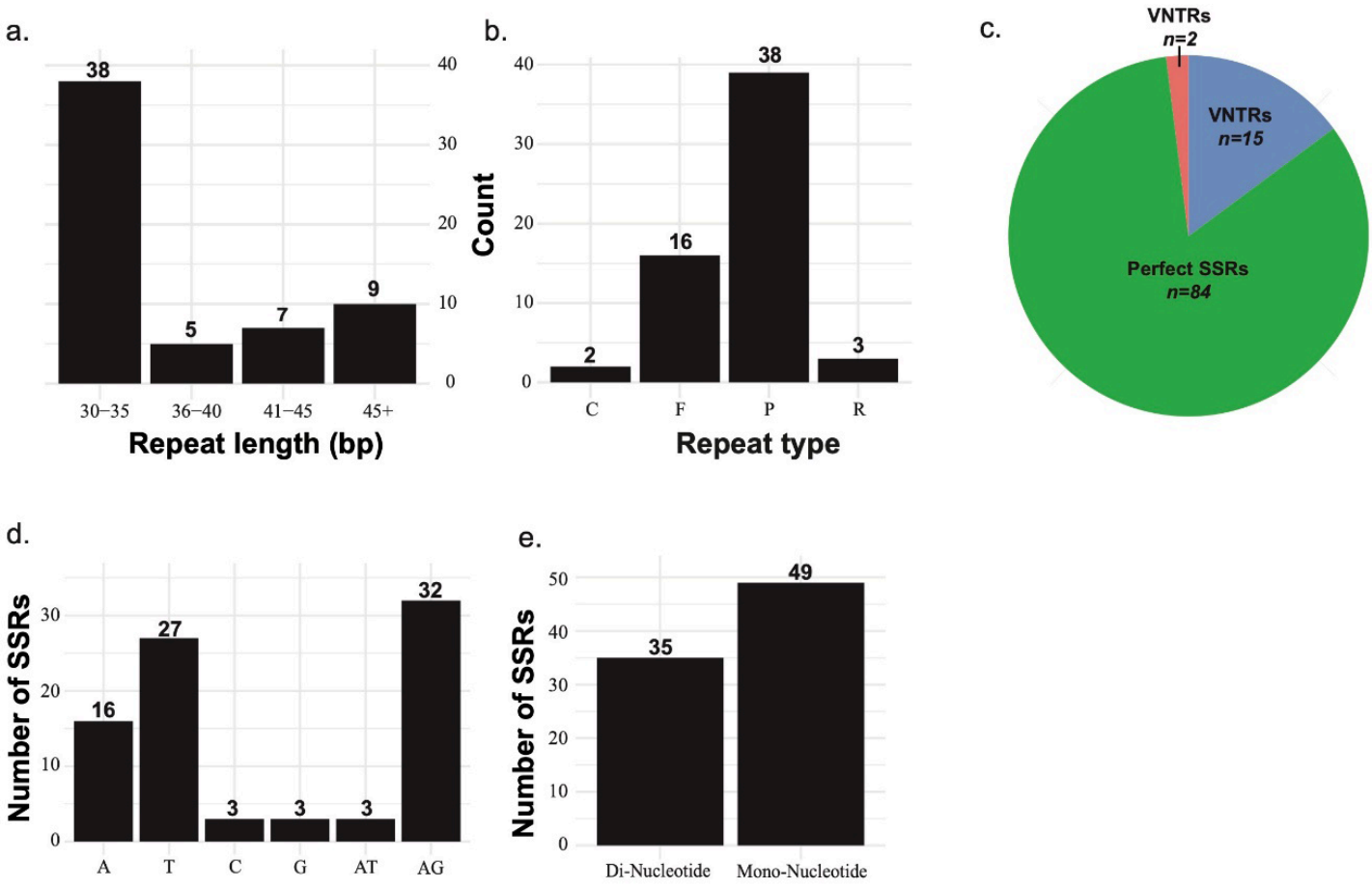


Figure 7

Repeat content and SSR analysis analysis in *P. tetragonolobus* chloroplast**.** a) Count distribution of repeats found in different length brackets b) Repeats count based on their type F (forward), R (reverse), P (palindromic) or C (complement) c) distribution of perfect and compound SSRs along with variable number tandem repeats (VNTRs) d-e) Difference in the presence of nucleotides defined in the SSRs.

Supplementary Files

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

- SupplementaryTableS1.docx
- SupplementaryTableS2.docx
- SupplementaryTableS3.docx
- SupplementaryTableS4.csv
- SupplementaryTableS5.docx