

Characterisation of the Chloroplast Genome of Macrotyloma Species: Comparative Analysis and Phylogenomic Insights

Ramawatar Nagar

nagarsaggi@gmail.com

ICAR-National Institute for Plant Biotechnology

Vandna Patial

ICAR-Indian Agricultural Research Institute

Divya Joshiya

ICAR-National Institute for Plant Biotechnology

Nimmy MS

ICAR-National Institute for Plant Biotechnology

Manavi Raizada

ICAR-National Institute for Plant Biotechnology

Pradeep Kumar Jain

ICAR-National Institute for Plant Biotechnology

Article

Keywords: Macrotyloma, chloroplast genome, phylogenomics, species divergence, species identification

Posted Date: February 10th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8730534/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on June 16th, 2026. See the published version at <https://doi.org/10.1038/s41598-026-55457-1>.

Abstract

Macrotyloma is an underutilised legume genus within the tribe Phaseoleae (Fabaceae) that includes nutritionally and agronomically important crops such as horse gram (*Macrotyloma uniflorum*) and Kersting's groundnut (*Macrotyloma geocarpum*). Despite their importance, knowledge of the chloroplast (cp) genome of this genus remains limited. In this study, we assembled and analysed the complete chloroplast genomes of three *Macrotyloma* species: *M. uniflorum*, *M. geocarpum*, and *M. axillare*. The chloroplast genomes were assembled into two isoforms that differ in the orientation of the small single-copy (SSC) region. Genome sizes ranged from 150,811 to 151,013 bp and exhibited the canonical quadripartite structure, comprising a pair of inverted repeats (IRa and IRb; 26,415 bp each), a large single-copy region (LSC; 80,446 bp), and a small single-copy region (SSC; 17,710 bp). Each genome encoded 126 genes, including 8 rRNA genes, 35 tRNA genes, and 83 protein-coding genes, and all three species possessed the papilionoid family-specific ~50 kb inversion. Although overall chloroplast genome structure and organisation were highly conserved among *Macrotyloma* species, gene-wise nucleotide diversity analysis identified seven relatively variable genes: *rps19*, *rps15*, *ccsA*, *ndhA*, *ycf1*, *ycf4*, and *psaI*. Phylogenomic analysis based on complete chloroplast genomes robustly resolved *Macrotyloma* as a monophyletic group within the Phaseolinae clade of the Papilionoideae subfamily. Within the genus, *M. uniflorum* and *M. axillare* formed a strongly supported sister pair, with *M. geocarpum* occupying a basal position. Overall, this study provides valuable insights into chloroplast genome evolution in *Macrotyloma* and enhances understanding of its phylogenetic placement within Phaseoleae, offering genomic resources for future evolutionary, taxonomic, and conservation studies of this underutilised legume genus.

Introduction

The genus *Macrotyloma* (Wight & Arn.) Verdc. belongs to the legume family Fabaceae, subfamily Papilionoideae, and tribe Phaseoleae¹. It comprises primarily African and Asian herbaceous or twining species, several of which are valued as forage crops and dietary protein sources for humans, particularly in arid and semi-arid regions². While many *Macrotyloma* species occur in the wild and serve as important reservoirs of genetic diversity, a few have been domesticated due to their ecological resilience and nutritional attributes. Among the cultivated species, *Macrotyloma uniflorum* (Lam.) Verdc., commonly known as horse gram, kulthi, or Madras gram, is a nutritionally rich yet underutilised pulse crop³. It is extensively cultivated in India, as well as in parts of Southeast Asia, Africa, and Australia⁴. *M. uniflorum* seeds are rich in protein, carbohydrates, dietary fibre, and essential minerals such as iron and calcium³. Another underutilised *Macrotyloma* species is the *Macrotyloma geocarpum* (Harms) Maréchal & Baudet, commonly known as Kersting's groundnut. It is grown and consumed in parts of West Africa⁵. *Macrotyloma axillare* (E. Mey.) Verdc. (Archer dal/dolichos) is a perennial forage legume valued in pastures for its high protein content and tolerance of heat and drought⁶. It is found to grow in both the Indian subcontinent and Africa. It is closely related to the *Macrotyloma uniflorum* and is believed to be the perennial progenitor of it⁷.

The chloroplast (cp) genome is a circular, highly conserved, and predominantly maternally inherited organellar genome in most angiosperms. It plays a central role in photosynthesis and in key metabolic pathways such as fatty acid and amino acid biosynthesis⁸. In most land plants, chloroplast genomes range from approximately 125 to 150 kilobase pairs (kb) and exhibit a conserved quadripartite structure consisting of a large single-copy (LSC) region and a small single-copy (SSC) region separated by two inverted repeats (IRs)^{9–14}. Variation in chloroplast genome size is generally limited and is primarily attributed to expansion and contraction of the IR regions^{15–17}.

Chloroplast DNA (cpDNA) has long been used in plant phylogenetics and taxonomy due to its haploid nature, uniparental inheritance, and relatively moderate evolutionary rate^{18–21}. Individual chloroplast genes such as *rbcL*, *matK*, and *trnL* have been widely employed in phylogenetic studies. For example, analyses combining *matK* sequences with morphological and chemical characters have contributed to the classification of Fabaceae into six subfamilies^{22,23}. However, reliance on a limited number of loci often lacks sufficient resolution to resolve complex evolutionary relationships, particularly among closely related tribes and genera.

Advances in high-throughput sequencing technologies have enabled the rapid generation and public availability of complete chloroplast genome sequences. This has facilitated whole-plastome-based phylogenomics, providing substantially improved resolution across multiple taxonomic levels^{24–26}. Comparative chloroplast genomics has further enabled detailed investigation of genome structure, rearrangements, IR boundary dynamics, mutational hotspots, and selective pressures, all of which contribute to a more robust understanding of plant evolutionary history²⁷.

In this study, we report the complete chloroplast genome sequences of three *Macrotyloma* species: the cultivated *M. uniflorum*, the underutilised *M. geocarpum*, and the perennial wild relative *M. axillare*. We characterise their genome structure, gene content, and repetitive elements, and perform comparative analyses to identify patterns of genomic divergence within the genus. Using whole chloroplast genome data, we further elucidate the phylogenetic placement and evolutionary history of *Macrotyloma* within the tribe Phaseoleae and across the Fabaceae family. Overall, this study advances our understanding of chloroplast genome evolution in *Macrotyloma* and provides valuable genomic resources for species identification, phylogenetic inference, germplasm conservation, and future utilisation of this underutilised legume genus.

Materials and methods

Plant samples, DNA extraction, and sequencing

Seeds of *Macrotyloma uniflorum*, commonly known as kulthi, were germinated under controlled laboratory conditions, and the seedlings were grown for two weeks. Total genomic DNA was extracted from the leaf tissue of a single healthy seedling using a modified cetyltrimethylammonium bromide (CTAB) protocol²⁸. High-quality total DNA was used for library preparation following the manufacturer's

instructions for the Nextera DNA Flex Library Prep Kit (Illumina, USA). Cluster generation and paired-end sequencing were performed on the Illumina NextSeq 2000 platform using 2 × 150 bp paired-end chemistry, in accordance with the manufacturer's protocols. The resulting paired-end reads were used for downstream genome assembly and analysis

Chloroplast genome assembly and annotation

The quality of the fastq files was assessed using the FastQC tool. To remove the adapter and poor-quality sequence, we use the Trim Galore v0.4.0. Chloroplast genome assemblies were generated using two independent tools, NOVOPlasty v2.7.2²⁹ and GetOrganelle v1.7.2³⁰. The Illumina sequence data were provided as input to the tools to assemble the chloroplast genome with default settings. The *de novo* assembly of the chloroplast genome by both tools relies on the mapping of whole genome reads to conserved chloroplast seed sequences, and then reads that are mapped to the seed are used as bait for the second iteration. The extension iteration continues until a chloroplast genome is assembled. To compare the chloroplast genome of the *Macrotyloma uniflorum* to other *Macrotyloma* species, the chloroplast genomes of the *Macrotyloma axillare* and *Macrotyloma geocarpum* were assembled using publicly available Illumina short-read sequencing data. For *Macrotyloma axillare*, raw sequencing reads were retrieved from accession SRR19031527 (BioProject PRJNA834023), and for *Macrotyloma geocarpum*, the raw reads were obtained from accession SRR31328150 (BioProject PRJNA1185675).

Genome annotation was performed using multiple chloroplast-specific annotation tools, including GeSeq³¹, CPGAVAS2³² and PGA³³. Annotated gene models were manually curated to confirm start codons, stop codons, and the absence of internal stop codons. Circular chloroplast genome maps were visualised using OGDRAW³⁴. The fully annotated chloroplast genome sequences were submitted to the NCBI GenBank database, and accession numbers were obtained. The Codon W in the Galaxy web server (<https://galaxy.pasteur.fr/?form=codonw>) was used to access the codon usage frequency and relative synonymous codon usage (RSCU) of protein-coding genes. To identify SSR repeats in the chloroplast genome of the *Macrotyloma* species, MISA-web was used³⁵. SSR were defined as repeating sequence sections of 1 to 5 bp that were repeated at least three times. The other repeats were identified using the Tendon repeat finder³⁶.

Comparative analysis of chloroplast genome structure

To investigate structural variation in the chloroplast genomes of *Macrotyloma* species, comparative analyses were conducted with representative members of the Phaseoleae tribe, *C. cajan* (KU729879), *Phaseolus lunatus* (NC_060403), *V. radiate* (NC_013843), *Sphenostylis stenocarpa* (PP920094), *Lablab purpureus* (NC_054310), three members of the IRLC, *G. max* (NC_007942), *C. arietinum* (NC_011163), and *M. sativa* (NC_032066), and *Arabidopsis thaliana* as a reference (NC_000932). Chloroplast genome sequences were retrieved from the NCBI database. All chloroplast genomes were reoriented with respect to the *rbcL* gene using the Rotate tool³⁷ to ensure consistency and subsequently reannotated using GeSeq³¹. Whole-genome alignments were performed using the progressiveMauve algorithm

implemented in Mauve to identify locally collinear blocks (LCB)³⁸. The plot identifies locally collinear blocks (LCBs), revealing structural conservation and variation across the included taxa

The boundaries of the large single-copy (LSC), small single-copy (SSC), and inverted repeat (IR) regions were visualised using CPJSDraw v1.0.0³⁹.

Nucleotide diversity analysis

To calculate gene-wise nucleotide diversity in the chloroplast genome of the *Macrotyloma* species, individual protein-coding genes, rRNA genes, and tRNA genes were extracted from the annotated GenBank files of the chloroplast genome using a custom Python script.

Each shared gene was individually aligned using MAFFT v7.525⁴⁰. The nucleotide diversity (π) of each gene was analysed separately using the DnaSP v6⁴¹. The π was estimated according to the method of Nei and Li⁴², which computes the mean number of pairwise nucleotide differences per site among all possible sequence pairs. The nucleotide of each gene was plotted using ggplot2 in the R environment.

Phylogenomics analysis

Phylogenetic relationships were inferred using chloroplast genome data from 33 species, including the three *Macrotyloma* species analysed in this study. Six species from the subfamily Caesalpinioideae were included as outgroups. Only RefSeq-validated chloroplast genomes were selected to ensure annotation reliability. Protein-coding sequences (CDS) shared across all species were extracted from GenBank files using a custom Python script. A total of 42 orthologous protein-coding genes were identified and concatenated in a consistent gene order for all taxa. The concatenated CDS alignments were generated using MAFFT v7.525 with the auto strategy⁴⁰. Poorly aligned and ambiguous regions were removed using trimAl v1.5.rev0 with automated trimming settings⁴³. Maximum-likelihood phylogenetic analysis was conducted using RAxML v8.2.12 (raxmlHPC-PTHREADS) under the GTRGAMMA nucleotide substitution model with 1,000 bootstrap replicates⁴⁴. The resulting phylogenetic tree was visualised and annotated in the R environment using the ggtree and ggplot2 packages^{45,46}.

Results

General features of the chloroplast genomes of *Macrotyloma* species

The chloroplast genome of all three *M. uniflorum* was assembled into two isoforms, which were varied for the orientation of the small single-copy region (Fig. 1). The presence of two isoforms, also known as structural heteroplasmy, has been widely reported among land plants. Similar to the *M. uniflorum* the two other species of *Macrotyloma* species, *Macrotyloma axillare*, *M. geocarpum*, also show structural heteroplasmy (Figure S1). The chloroplast genome of all three species showed a highly conserved quadripartite structure, comprising a large single-copy (LSC) region, a small single-copy (SSC) region, and two copies of the inverted repeats (IRs) (Fig. 1 and Table 1). Genome sizes were comparable among

the species, ranging from 150,811 bp in *M. geocarpum* to 151,013 bp in *M. uniflorum*, indicating minimal structural divergence across the genus. The LSC regions ranged from 80,229 to 80,446 bp, while the SSC region was nearly identical across species (17,710 – 17,711 bp). The IRs were also highly similar in length (52,830 – 52,870 bp), suggesting stable expansion-contraction dynamics among the three chloroplasts. Gene content was also highly consistent across the *Macrotyloma*, including 83 protein-coding genes, 8 rRNA genes and 35 tRNAs (Table 1). Coding regions constituted approximately 69.61–70.61% of the total genome. Non-coding sequence length ranged from 44,321 bp to 45,891 bp and was most abundant in *M. uniflorum*. The overall GC content was also highly consistent, ranging from 35.3% to 35.4%. Gene density was uniform (0.83–0.84 genes per kb), suggesting similar levels of coding compactness among species.

Table 1
General features of the chloroplast genome of three *Macrotyloma* species

Feature	<i>Macrotyloma axillare</i>	<i>Macrotyloma geocarpum</i>	<i>Macrotyloma uniflorum</i>
GenBank Accession No.	PX659840	PX659841	PX549914
Total chloroplast genome size (bp)	150989	150811	151013
LSC length (bp)	80446	80229	80437
SSC length (bp)	17711	17710	17710
IR length (bp, both, IRa and IRb)	52830	52870	52864
Total number of genes	126	127	125
Protein-coding genes	83	83	83
tRNA genes	35	35	35
rRNA genes	8	8	8
Length of protein-coding genes (bp)	87430	87434	86821
Length of tRNA genes (bp)	9225	9980	9225
Length of rRNA genes (bp)	9076	9076	9076
Total coding region length (bp)	105731	106490	105122
Non-coding region length (bp)	45258	44321	45891
Protein-coding proportion (%)	57.9	57.98	57.49
tRNA proportion (%)	6.11	6.62	6.11
rRNA proportion (%)	6.01	6.02	6.01
Total coding region proportion (%)	70.03	70.61	69.61
Non-coding region proportion (%)	29.97	29.39	30.39
Total GC content (%)	35.3	35.4	35.3
Gene density (genes/kb)	0.83	0.84	0.83

Gene Content and Functional Categorisation

The chloroplast genomes of the *M. uniflorum*, along with the sister species, comprise a complete set of ATP synthase, photosystem I and II, cytochrome b_6/f complex, rubisco (*rbcL*), and NADH dehydrogenase (*ndh*) subunits (Table 2). Genes involved in self-replication included multiple ribosomal protein genes associated with both the small (SSU) and large (LSU) ribosomal subunits, as well as the four RNA polymerase subunits (*rpoA-rpoC2*). Several other essential genes, such as *accD*, *ccsA*, *cemA*, *clpP*, and *matK*, were also identified. Along with functional genes, the chloroplasts retained a set of conserved open reading frames (*ycf1*, *ycf2*, *ycf3*, and *ycf4*). These include ribosomal protein genes (*rpl2*, *rpl23*, *rps7*, *rps12*), the NADH dehydrogenase subunit *ndhB*, multiple tRNA genes (e.g., *trnA*-UGC, *trnI*-CAU, *trnN*-GUU, *trnR*-ACG, *trnV*-GAC), and all four rRNA genes (*rrn16*, *rrn23*, *rrn4.5*, *rrn5*). The duplication of these genes contributes to increased gene dosage and is thought to enhance the structural stability of the chloroplast genome. All the protein-coding genes in the chloroplast genome of *Macrotyloma* species start with a canonical start codon (ATG), except for the *ndhD* and *psbL* (ACG). Many genes are identified to carry introns. A total of 20 intron-containing genes were identified (four are duplicated in the inverted repeat), including 13 protein-coding genes and 7 tRNA genes (Table 3). Among these, two protein-coding genes, *clpP* and *ycf3*, contained two introns. Several intron-containing genes were located within the inverted repeat regions. The genes *rpl2*, *ndhB*, *trnA*-UGC, and *trnE*-UUC were each present in two copies, corresponding to the IRa and IRb. The largest intron was observed in *trnK*-UUU (2,560 bp), which harbours the *matK* gene.

Table 2
Gene composition of the chloroplast genome of *Macrotyloma* species

Function	Gene group	Gene
Biosynthesis of fatty acids	Acetyl-CoA carboxylase	<i>accD</i>
Genetic apparatus	Large subunit of ribosomal protein	rpl2 , <i>rpl14</i> , <i>rpl16</i> , <i>rpl20</i> , rpl23 , <i>rpl32</i> , <i>rpl33</i> , <i>rpl36</i>
	Small subunit of ribosomal protein	<i>rps2</i> , <i>rps3</i> , <i>rps4</i> , rps7 , <i>rps8</i> , <i>rps11</i> , rps12 , <i>rps14</i> , <i>rps15</i> , <i>rps16</i> , <i>rps18</i> , <i>rps19</i>
	Subunits of RNA polymerase	<i>rpoA</i> , <i>rpoB</i> , <i>rpoC1</i> , <i>rpoC2</i>
Photosynthesis pathway	Photosystem I	<i>psaA</i> , <i>psaB</i> , <i>psaI</i> , <i>psaJ</i>
	Photosystem I assembly	<i>ycf3</i> , <i>ycf4</i>
	Photosystem II	<i>psbA</i> , <i>psbB</i> , <i>psbC</i> , <i>psbD</i> , <i>psbE</i> , <i>psbF</i> , <i>psbH</i> , <i>psbI</i> , <i>psbJ</i> , <i>psbK</i> , <i>psbL</i> , <i>psbM</i> , <i>psbN</i> , <i>psbT</i> , <i>psbZ</i>
	F-type ATP synthase	<i>atpA</i> , <i>atpB</i> , <i>atpE</i> , <i>atpF</i> , <i>atpH</i> , <i>atpI</i>
	NADH-plastoquinone oxidoreductase	<i>ndhA</i> , ndhB , <i>ndhC</i> , <i>ndhD</i> , <i>ndhE</i> , <i>ndhF</i> , <i>ndhG</i> , <i>ndhH</i> , <i>ndhI</i> , <i>ndhJ</i> , <i>ndhK</i>
	Component of cytochrome b6/f complex	<i>petA</i> , petB , petD , <i>petG</i> , <i>petL</i> , <i>petN</i>
	Carbon metabolism	<i>cemA</i>
	Cytochrome-c biogenesis protein	<i>ccsA</i>
	Large subunit of Rubisco	<i>rbcL</i>
Protein-modifying	ATP-dependent protease proteolytic subunit	<i>clpP</i>
Structural RNAs	Transfer RNAs	trnA-UGC , <i>trnC-GCA</i> , <i>trnD-GUC</i> , trnE-UUC , <i>trnF-GAA</i> , <i>trnG-GCC</i> , <i>trnG-UCC</i> , <i>trnH-GUG</i> , trnI-CAU , <i>trnI-GAU</i> , <i>trnK-UUU</i> , trnL-CAA , <i>trnL-UAA</i> , <i>trnL-UAG</i> , <i>trnM-CAU</i> , trnN-GUU , <i>trnP-UGG</i> , <i>trnQ-UUG</i> , trnR-ACG , <i>trnR-UCU</i> , <i>trnS-GCU</i> , <i>trnS-GGA</i> , <i>trnS-UGA</i> , <i>trnT-GGU</i> , <i>trnT-UGU</i> , trnV-GAC , <i>trnV-UAC</i> , <i>trnW-CCA</i> , <i>trnY-GUA</i> , <i>trnM-CAU</i>
	Ribosomal RNAs	rrn16 , rrn23 , rrn4.5 , rrn5

Function	Gene group	Gene
Post-transcriptional modification	Maturase	<i>matK</i>
Protein of unknown function	Other genes	<i>ycf1, ycf2</i>
☒ <i>rp12</i> contains an intron and is duplicated in the IR regions. ☒ Gene contains one or more introns. ☒ <i>rps12</i> is a trans-spliced gene, with the 5' exon located in the LSC region and the duplicated 3' exons located in the IR regions.		

Table 3
Lengths of introns and exons for intron-containing genes in the chloroplast genome of *Macrotyloma geocarpum*

Gene	Strand	Exon I (bp)	Intron I (bp)	Exon II (bp)	Intron II (bp)	Exon III (bp)
clpP	+	71	715	294	768	226
ycf4	–	175	30	410	–	–
accD	–	99	93	1233	–	–
trnM-CAU	+	42	657	57	–	–
atpF	–	145	704	398	–	–
rpoC1	–	432	825	1623	–	–
ycf3	–	124	694	230	782	153
trnL-UAA	+	36	511	51	–	–
trnC-GCA	–	39	561	56	–	–
trnK-UUU	+	37	2560	35	–	–
rpl2 _{IRa}	–	391	655	434	–	–
ndhB _{IRa}	–	721	691	758	–	–
trnE-UUC _{IRa}	+	32	950	40	–	–
trnA-UGC _{IRa}	+	37	803	36	–	–
ycf1	+	4374	30	960	–	–
ndhA	+	553	1257	539	–	–
trnA-UGC _{IRb}	–	37	803	36	–	–
trnE-UUC _{IRb}	–	32	950	40	–	–
ndhB _{IRb}	+	721	691	758	–	–
rpl2 _{IRb}	+	391	655	434	–	–
_{IRa} Gene copy located in IRa _{IRb} Gene copy located in IRb						

SSR contents and composition

The chloroplast genome of the *Macrotyloma* species contained 58 to 66 SSRs (CpSSR) (Table S4-6). In all cases, the intergenic spacer (IGS) regions harboured the highest number of SSRs (41–44). Gene regions contained a moderate number of SSRs (11–12), while intronic sequences showed the lowest SSR density (5–9) (Fig. 4). Among the SSR repeats, the mononucleotide repeats were the highest

(Fig. 2A-B). The A-rich and T-rich motifs accounted for the majority of SSRs. The dinucleotide motifs (AT and TA) occurred less frequently (144 occurrences), whereas other motif categories were rare or nearly absent. This predominance of A/T mononucleotide repeats is consistent with the AT-rich nature of angiosperm chloroplast genomes and aligns with previous reports on plastid SSR composition.

Codon Uses Analysis

The total number of codons found in the plastid genome of the *Macrotyloma* species varied from 20587 to 20595 (Table S 1–3). The codon usage analysis of the *Macrotyloma* species revealed that the codons TTT (Phe), ATT (Ile), TTA (Leu), GAT (Asp), GAA (Glu), and TAT (Tyr) exhibited higher usage frequencies, consistent with an overall preference for A/T-rich codons (Table). This trend aligns with the AT-biased nucleotide composition characteristic of chloroplast genomes⁴⁷. Conversely, codons such as ATC (Ile) and TGT (Cys) were used less frequently, reflecting selective constraints on synonymous substitution and codon choice. Overall, all genomes show a strong AT bias, particularly at the third codon position; TTT (Phe) and ATT (Ile) are the most frequently used codons; leucine is the most abundant amino acid, cysteine is the least abundant, and TAA is the preferred stop codon.

Comparative Genome Structure Analysis

To explore large-scale genome rearrangements and highlight lineage-specific structural changes, we compared the chloroplast genomes of *Macrotyloma* species, eight other legumes, and *Arabidopsis thaliana* as a non-legume reference using the Mauve software (Fig. 3). The Cp genomes of all three *Macrotyloma* species (*M. axillare*, *M. geocarpum*, *M. uniflorum*) are highly conserved and collinear, with no major rearrangements. This syntenic organisation is also shared with close relatives such as *Sphenostylis stenocarpa* and *Lablab purpureus*. However, when compared to *A. thaliana* and other legume species, it revealed several distinct features. All the leagues, including *Macrotyloma* species, carry the characteristic Papilionoideae subfamily-specific 50-kb inversion in the LSC region. The size of the inversion varies slightly between the *Macrotyloma* species. The *M. uniflorum* has a slightly bigger inversion (~ 500 bp) compared to the two other species. As expected, members of the IR-lacking clade (IRLC), *G. max*, *C. arietinum*, and *M. sativa* have lost one inverted repeat and exhibit substantial rearrangements and inversion events relative to both *Macrotyloma* and *Arabidopsis*.

Inverted Repeats Contraction and Expansion Analysis

The boundaries and sizes of the different sub-structures of the chloroplast genome were also compared among *Macrotyloma* species and representative members of the Phaseoleae tribe (Fig. 4). Across *Macrotyloma*, the lengths of the LSC (70,543 – 70,784 bp), SSC (17,710 – 17,711 bp), and IR regions (~ 26,416 – 26,436 bp) were highly conserved. The IRb-LSC junction (JLB) consistently intersected the *rps19* gene, resulting in partial duplication within the IR. This configuration was shared with *Phaseolus* and *Sphenostylis*, whereas *Vigna radiata* showed a slightly reduced *rps19* extension, indicating minor IR contraction. At the IRb-SSC junction (JSB), *Macrotyloma* chloroplasts uniformly place *ycf1* across the boundary, with ~ 474–528 bp extending into the IRb. Although a similar arrangement was observed in

other Phaseoleae species, *Phaseolus vulgaris* displayed a larger duplicated portion of *ycf1* (~ 616 bp), displaying a modest IR expansion. The SSC-IRa junction (JSA) was also conserved across *Macrotyloma*, with *ndhF* positioned adjacent to the boundary and minimal overlap (1–18 bp) into the IRa; slightly greater *ndhF* extensions were observed in *Sphenostylis* and *Vigna*. At the IRa-LSC junction (JLA), partial overlap of *rps19* occurred in all *Macrotyloma* species, mirroring patterns in *Phaseolus* and *Sphenostylis*, though with reduced overlap in *Sphenostylis* and *Vigna*.

Nucleotide diversity analysis

Although the chloroplast genomes of *Macrotyloma* species are highly conserved in structure and gene content organisation, nucleotide diversity analysis revealed pronounced heterogeneity in evolutionary rates across genes (Fig. 5). Most chloroplast genes exhibited very low nucleotide diversity, reflecting strong purifying selection. However, the chloroplast genome of the *Macrotyloma* seven genes, including *rps19*, *rps15*, *ccsA*, *ndhA*, *ycf1*, *ycf4* and *psal*, showed markedly elevated nucleotide diversity, indicating relaxed selective constraints or lineage-specific divergence. Among these, *psal*, a small, hydrophobic subunit of Photosystem I (PSI), is the most variable gene. The *psal*, along with the *accD*, often hyphenate together as *accD-psal*, has been identified as a hotspot of genetic diversity in the plant chloroplast genome. The gene, *ycf1*, has been reported to be one of the most variable plastid genes in angiosperms and displayed high diversity in *Macrotyloma*, reinforcing its utility as a molecular marker for species discrimination and for population-level studies. These findings demonstrate that, despite overall plastome conservation, selective pressures vary among genes, providing insight into chloroplast genome evolution and identifying candidate loci for future evolutionary and functional analyses.

Phylogenetic analysis

To elucidate the evolutionary relationship of *Macrotyloma* species within the Fabaceae family, we reconstructed a maximum-likelihood (ML) phylogeny based on the concatenated coding sequences (CDS) of 42 shared chloroplast genes. The analysis included 34 representative taxa, *M. uniflorum*, *M. axillare*, and *M. geocarpum*, along with 31 additional species spanning the major clades within the Papilionoideae subfamily. For the outgroup, five species from the Caesalpinioideae subfamily, representing a lineage outside of Papilionoideae, were selected to root the tree. The resulting phylogeny robustly placed the *Macrotyloma* species within the Phaseolinae clade of Papilionoideae subfamily (Fig. 6). Within this group, the three *Macrotyloma* species formed a strongly supported monophyletic subclade (bootstrap = 100%), confirming their close evolutionary relationship. *M. axillare* and *M. uniflorum* were resolved as sister taxa, with *M. geocarpum* branching as a nearby lineage, consistent with the hypothesis that *M. axillare* is the wild progenitor of cultivated *M. uniflorum*. The closest relative to the *Macrotyloma* subclade was *Sphenostylis stenocarpa* (African yam bean), and *Lablab purpureus*, both of which are legumes of African origin. This relationship underscores the biogeographic affinity of *Macrotyloma* to other indigenous African legumes and may provide insights into the domestication and diversification history of *M. geocarpum* and *M. uniflorum* across African and Asian centres of

cultivation. The analysis supports the utility of chloroplast sequences in resolving deep and shallow phylogenetic relationships within legumes.

Discussion

Characteristics of the chloroplast genome of *Macrotyloma* species.

The size of the chloroplast genome of the *Macrotyloma* species varied 150,811 – 151,013 bp, a size range reported for most members of the Phaseolus tribe. For example, *Vigna radiata* (151,024 bp; Tangphatsornruang et al., 2010), *Phaseolus vulgaris* (150,785 bp; Meng and Li, 2018), *Glycine max* (152,218 bp; Saski et al., 2005), and *Cajanus cajan* (152,224 bp; Kaila et al., 2016). The relatively small variation in genome size among *Macrotyloma* species primarily arises from minor differences in the length of the IR regions, a feature known to drive chloroplast size polymorphism in the chloroplast genome^{49–52}. The overall GC content (35.3–35.4%) is nearly identical among the three *Macrotyloma* species and consistent with values reported for other legumes of the Phaseolus tribe^{10,14,53}. The GC content distribution across the LSC (32.7–32.8%), SSC (28.6–28.7%), and IR (35.0%), where IR regions are GC-rich due to the presence of rRNA and tRNA genes^{54,54,55}. Each chloroplast encodes 126 genes, including 83 protein-coding genes, 35 tRNA genes, and 8 rRNA genes. The number and density genes were similar to those of the close member, *Glycine max* (127 genes; Saski et al., 2005), *Cajanus cajan* (127 genes; Kaila et al., 2017, 2016), and *Phaseolus vulgaris* (126 genes; Meng and Li, 2018). All the protein-coding genes use the canonical start codon, ATG, except for *ndhD* and *psbL*, which use the alternative start codon ACG. The start codon of these genes is post-transcriptionally converted to the canonical AUG start codon through C-to-U RNA editing^{57–59}. A total of 20 intron-containing genes were identified, including *clpP* and *ycf3*, each harbouring two introns. The largest intron occurred in *trnK-UUU*, which contains the *matK* gene, as reported in many plant species⁶⁰. The trans-splicing of *rps12* further reflects a conserved chloroplast gene architecture shared across land plants⁶¹.

We also analysed repetitive sequences in the chloroplast genomes of *Macrotyloma* species, which play important roles in genome rearrangements and variation within intergenic regions. The *Macrotyloma* chloroplast genomes contained 56–64 chloroplast simple sequence repeats (cpSSRs). The abundance of cpSSRs in *Macrotyloma* was relatively low compared with other members of the Fabaceae family, such as lima bean (cpSSRs: 290), pigeonpea (cpSSRs: 290), and cluster bean (cpSSRs: 220)^{9,56,62}. Chloroplast SSRs are considered valuable molecular markers for evolutionary and taxonomic studies due to their maternal inheritance and relatively low mutation rates^{63,64}. Consistent with observations in other angiosperms, the cpSSRs identified in *Macrotyloma* were predominantly A/T-rich and located primarily in intergenic regions. These cpSSRs represent informative, non-destructive genetic markers capable of capturing fine-scale genetic variation and therefore hold considerable potential for population genetic and phylogeographic studies in this underexplored genus⁶⁵.

Structural variation and IR Contraction/Expansion

The organisation of the chloroplast genome is largely conserved across plant species⁶⁶. However, inversions, rearrangements, and expansions or contractions of the inverted repeat (IR) regions have been frequently reported⁶⁶. Within the Fabaceae, several lineage-defining structural variations have been documented⁶⁷. For example, a large inversion of approximately 50–70 kb LSC region is characteristic of the Papilionoideae subfamily⁶⁸. Furthermore, within this Papilionoideae subfamily, a monophyletic group of legumes has completely lost one copy of the inverted repeat, a feature that defines the inverted repeat-lacking clade (IRLC)^{65,67}. The chloroplast genomes of all three *Macrotyloma* species were assembled into two isoforms that differed in the orientation of the small single-copy (SSC) region (Fig. 1). The presence of the two structural isoforms is referred to as structural heteroplasmy, which is caused by flip-flop recombination⁶⁹. Except for the structural heteroplasmy, the plastomes of *M. axillare*, *M. geocarpum*, and *M. uniflorum* are highly conserved and collinear. However, when compared to other legume species and nanolegume species, the plastid of the *Macrotyloma* species has retained the 50 kb inversion in the LSC spanning from *trnK* to *accD*, a synapomorphy feature of legumes^{68,70}.

Nucleotide diversity

Chloroplast genomes are generally subject to strong purifying selection, which limits sequence variation across most protein-coding genes⁷¹. Nevertheless, certain chloroplast genes consistently emerge as hypervariable regions across angiosperms and legumes⁶². Gene-wise nucleotide diversity (π) analysis revealed moderate to elevated diversity in several coding regions, including *rps19*, *rps15*, *ccsA*, *ndhA*, *ycf1*, *ycf4*, and *psaI* (Figure). Among these, *ycf1* is widely recognised as one of the most variable chloroplast coding genes^{72,73}. In *Macrotyloma*, *ycf1* also show high nucleotide diversity, consistent with its role as a plastome diversity hotspot reported in other legume lineages⁷³. Adjacent to *ycf1*, the ribosomal protein gene *rps15* also showed elevated diversity. Although *rps15* is functionally conserved, it is frequently located within a highly variable genomic context, particularly the *rps15-ycf1* intergenic region, which is prone to mutation accumulation^{74,75}. The *psaI* gene, typically part of the *accD-psaI-ycf4* gene block, also displayed increased nucleotide diversity⁷⁶. The moderate diversity observed in *ycf4*, *ccsA*, and *ndhA* further supports the notion that ribosomal protein genes and photosynthesis-related genes can accumulate lineage-specific variation despite overall plastome conservation. These hypervariable regions represent promising molecular markers for resolving phylogenetic relationships within *Macrotyloma* and across the Leguminosae. In many previous studies, genes such as *ycf1*, *accD*, *ndhF*, and noncoding regions including *trnL-trnF* and *atpB-rbcL* have been used to infer evolutionary relationships, domestication histories^{77,78}, and centres of origin in legume crops such as *Phaseolus lunatus* and *Quercus bawanglingensis*^{79,80}. The gene-wise nucleotide diversity patterns observed here highlight the value of expanding plastome resources for underutilised legumes and provide a foundation for future population genetic, phylogeographic, and domestication studies in *Macrotyloma*.

Phylogenetic Relationships of *Macrotyloma* within Fabaceae

So far, very limited molecular data have been used to infer the phylogeny of the *Macrotyloma* species⁸¹. The availability of complete chloroplast genomes has significantly enhanced phylogenetic resolution in plant systematics⁸². In this study, we reconstructed a maximum-likelihood phylogeny based on 42 shared chloroplast coding genes across 34 Fabaceae species, including three *Macrotyloma* species, to infer their intra- and intergeneric relationships (Figure). The phylogeny robustly places *Macrotyloma* within the Phaseolinae clade of the Papilionoideae subfamily^{83,84}. All three species, *M. uniflorum*, *M. axillare*, and *M. geocarpum*, form a well-supported monophyletic group. With the group, *M. uniflorum* and *M. axillare* appear as sister taxa, consistent with hypotheses that *M. axillare* is a perennial progenitor of the domesticated *M. uniflorum*^{7,81}. The African species *M. geocarpum* clusters close to this pair, forming a cohesive lineage distinct from other Phaseoleae genera. It suggests that the genus's cultivated members were derived from the same ancestral gene pool. The horsegram (*M. uniflorum*) is widely grown in India and other parts of Asia⁸⁵, whereas Kersting's groundnut (*M. geocarpum*) is a traditional legume of West Africa⁸⁶. The divergence between Asian (*M. uniflorum*) and African (*M. geocarpum*) lineages suggests independent domestication events following an early geographic split. In summary, our phylogenomic analysis affirms the monophyly and distinct evolutionary trajectory of *Macrotyloma* within the Phaseoleae tribe. This placement refines our understanding of its domestication history and supports targeted efforts for genetic improvement and conservation of these underutilised legumes^{23,84,87}

Conclusions

This study presents the first comprehensive comparative analysis of complete chloroplast genomes from multiple *Macrotyloma* species, providing new genomic resources for an underutilised but agriculturally important legume genus. Although the chloroplast genome of *Macrotyloma* species exhibits a high degree of structural conservation, gene content and organisation, they exhibit a moderate level of nucleotide diversity among genes. A few genes stand out as highly variable, such as *rps19*, *rps15*, *ccsA*, *ndhA*, *ycf1*, *ycf4* and *psaI*, which can be used for barcoding and species identification markers. Chloroplast phylogenomic analyses robustly resolve *Macrotyloma* as a monophyletic lineage within the Phaseolinae clade of Papilionoideae and clarify its relationships with other Phaseoleae genera. The close sister relationship between *M. uniflorum* and *M. axillare* is consistent with previous hypotheses of a close evolutionary affinity. At the same time, the basal placement of *M. geocarpum* highlights early divergence within the genus. Together, these findings refine our understanding of *Macrotyloma* evolution and its position within legume phylogeny. Beyond evolutionary insights, the annotated chloroplast genomes and identified cpSSRs reported here provide valuable tools for future studies on population genetics, phylogeography, and conservation. Given the importance of *Macrotyloma* species for climate-resilient agriculture and food security, the chloroplast genome resources of *Macrotyloma* species offer a genomic foundation to support germplasm conservation and the genetic improvement of these lesser-known legume crops.

Declarations

Competing interests

The author declares no competing interests.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author Contribution

RN conceived the study, designed the experiments, and carried them out. RN performed all bioinformatics analyses, including chloroplast genome assembly and annotation, and wrote the manuscript. VP, DJ, and MR contributed to DNA extraction, quality control analyses, and manuscript preparation. PKJ and NMS provided critical suggestions and revisions to the manuscript.

Data Availability

The assembled chloroplast genomes have been submitted to the NCBI with the accession numbers PX659840, PX659841 and PX549914.

References

1. LACKEY, J. A. A revised classification of the tribe Phaseoleae (Leguminosae: Papilionoideae), and its relation to canavanine distribution. *Bot. J. Linn. Soc.* **74**, 163–178 (1977).
2. Verdcourt, B. Studies in the Leguminosae-Papilionoideae for the 'Flora of Tropical East Africa': IV. *Kew Bull.* **24**, 507–569 (1970).
3. Prasad, S. K. & Singh, M. K. Horse gram- an underutilized nutraceutical pulse crop: a review. *J. Food Sci. Technol.* **52**, 2489–2499 (2015).

4. Fuller, D. Q. & Murphy, C. The origins and early dispersal of horsegram (*Macrotyloma uniflorum*), a major crop of ancient India. *Genet. Resour. Crop Evol.* **65**, 285–305 (2018).
5. Cheung, T. Y., Kafoutchoni, K. M., Agoyi, E. E., Chan, T. F. & Chapman, M. A. Exceptionally low genomic diversity in the underutilised legume Kersting's groundnut. *Nat. Commun.* **16**, 5183 (2025).
6. Fisher, D., Reynolds, I. & Chapman, M. A. The Perennial Horse Gram (*Macrotyloma axillare*) Genome, Phylogeny, and Selection Across the Fabaceae. in *Underutilised Crop Genomes* (ed Chapman, M. A.) 255–279 (Springer International Publishing, Cham, doi:10.1007/978-3-031-00848-1_14. (2022).
7. Morris, J. B. *Macrotyloma axillare* and *M. uniflorum*: descriptor analysis, anthocyanin indexes, and potential uses. *Genet. Resour. Crop Evol.* **55**, 5–8 (2008).
8. Dobrogojski, J., Adamiec, M. & Luciński, R. The chloroplast genome: a review. *Acta Physiol. Plant.* **42**, 98 (2020).
9. Kaila, T. et al. Chloroplast Genome Sequence of Pigeonpea (*Cajanus cajan* (L.) Millspaugh) and *Cajanus scarabaeoides* (L.) Thouars: Genome Organization and Comparison with Other Legumes. *Front. Plant. Sci.* **7**, 1847 (2016).
10. Kato, T., Kaneko, T., Sato, S., Nakamura, Y. & Tabata, S. Complete structure of the chloroplast genome of a legume, *Lotus japonicus*. *DNA Res. Int. J. Rapid Publ Rep. Genes Genomes.* **7**, 323–330 (2000).
11. Li, J. et al. Comparative analysis of the chloroplast genomes of eight *Piper* species and insights into the utilization of structural variation in phylogenetic analysis. *Front. Genet.* **13**, 925252 (2022).
12. Li, L. et al. Comparative analysis of 18 chloroplast genomes reveals genomic diversity and evolutionary dynamics in subtribe Malaxidinae (Orchidaceae). *BMC Plant. Biol.* **25**, 1013 (2025).
13. Saski, C. et al. Complete chloroplast genome sequence of *Glycine max* and comparative analyses with other legume genomes. *Plant. Mol. Biol.* **59**, 309–322 (2005).
14. Tangphatsornruang, S. et al. The Chloroplast Genome Sequence of Mungbean (*Vigna radiata*) Determined by High-throughput Pyrosequencing: Structural Organization and Phylogenetic Relationships. *DNA Res. Int. J. Rapid Publ Rep. Genes Genomes.* **17**, 11–22 (2010).
15. Chang, C. C. et al. The Chloroplast Genome of *Phalaenopsis aphrodite* (Orchidaceae): Comparative Analysis of Evolutionary Rate with that of Grasses and Its Phylogenetic Implications. *Mol. Biol. Evol.* **23**, 279–291 (2006).
16. Yue, F., Cui, L., dePamphilis, C. W., Moret, B. M. & Tang, J. Gene rearrangement analysis and ancestral order inference from chloroplast genomes with inverted repeat. *BMC Genom.* **9**, S25 (2008).
17. Aldrich, J., Cherney, B. W. & Williams, C. Sequence analysis of the junction of the large single copy region and the large inverted repeat in the petunia chloroplast genome. *Curr. Genet.* **14**, 487–492 (1988).
18. Daniell, H., Lin, C. S., Yu, M. & Chang, W. J. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biol.* **17**, 134 (2016).

19. Guo, C. et al. Maternal donor and genetic variation of *Lagerstroemia indica* cultivars. *Int. J. Mol. Sci.* **24**, 3606 (2023).
20. Li, E. et al. Insights into the phylogeny and chloroplast genome evolution of *Eriocaulon* (*Eriocaulaceae*). *BMC Plant. Biol.* **23**, 32 (2023).
21. Sablok, G., Mudunuri, S. B., Edwards, D. & Ralph, P. J. Chloroplast genomics: Expanding resources for an evolutionary conserved miniature molecule with enigmatic applications. *Curr. Plant. Biol.* **7–8**, 34–38 (2016).
22. 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).
23. Azani, N. et al. A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny: The Legume Phylogeny Working Group (LPWG). *TAXON* **66**, 44–77 (2017).
24. Dong, W. et al. Phylogenomic approaches untangle early divergences and complex diversifications of the olive plant family. *BMC Biol.* **20**, 92 (2022).
25. Alzahrani, D. A., Yaradua, S. S., Albokhari, E. J. & Abba, A. Complete chloroplast genome sequence of *Barleria prionitis*, comparative chloroplast genomics and phylogenetic relationships among *Acanthoideae*. *BMC Genom.* **21**, 393 (2020).
26. Lemieux, C., Otis, C. & Turmel, M. Comparative Chloroplast Genome Analyses of Streptophyte Green Algae Uncover Major Structural Alterations in the Klebsormidiophyceae, Coleochaetophyceae and Zygnematophyceae. *Front Plant. Sci* **7**, (2016).
27. Sabir, J. S. M. et al. Conserved Gene Order and Expanded Inverted Repeats Characterize Plastid Genomes of Thalassiosirales. *PLOS ONE*. **9**, e107854 (2014).
28. Healey, A., Furtado, A., Cooper, T. & Henry, R. J. Protocol: a simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species. *Plant. Methods*. **10**, 21 (2014).
29. Dierckxsens, N., Mardulyn, P. & Smits, G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* **45**, e18 (2017).
30. Jin, J. J. et al. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* **21**, 241 (2020).
31. Tillich, M. et al. GeSeq - versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* **45**, W6–W11 (2017).
32. Shi, L. et al. CPGAVAS2, an integrated plastome sequence annotator and analyzer. *Nucleic Acids Res.* **47**, W65–W73 (2019).
33. Qu, X. J., Moore, M. J., Li, D. Z. & Yi, T. S. PGA: a software package for rapid, accurate, and flexible batch annotation of plastomes. *Plant. Methods*. **15**, 50 (2019).

34. Greiner, S., Lehwark, P. & Bock, R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* **47**, W59–W64 (2019).
35. Beier, S., Thiel, T., Münch, T., Scholz, U. & Mascher, M. MISA-web: a web server for microsatellite prediction. *Bioinformatics* **33**, 2583–2585 (2017).
36. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
37. Durbin, R., De Sanctis, B., Blumer, M. & Rotate A command-line program to rotate circular DNA sequences to start at a given position or string. *Wellcome Open. Res.* **8**, 401 (2023).
38. Darling, A. C. E., Mau, B., Blattner, F. R., Perna, N. T. & Mauve Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Res.* **14**, 1394–1403 (2004).
39. Huang, L., Yu, H., Wang, Z. & Xu, W. CPStools: A package for analyzing chloroplast genome sequences. *iMetaOmics* **1**, e25 (2024).
40. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
41. Rozas, J. et al. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Mol. Biol. Evol.* **34**, 3299–3302 (2017).
42. Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. U. S. A.* **76**, 5269–5273 (1979).
43. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
44. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
45. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis* (Springer Publishing Company, 2009).
46. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
47. Fang, J., Qin, L., Liu, H. & Hu, Z. Analysis of Codon Usage Bias of 30 Chloroplast Genomes in Ulva (Ulvophyceae, Chlorophyta). *Genes* **16**, 608 (2025).
48. Meng, X. & Li, M. Characterisation of the complete chloroplast genome of the common bean, *Phaseolus vulgaris* L. *Mitochondrial DNA Part. B.* **3**, 918–920 (2018).
49. Abdullah et al. Inverted repeats dynamics shape Asclepiadoideae (Apocynaceae) Chloroplast genomes: effects on genome size, gene content, structural arrangement, and mutation rate. *BMC Genom.* **26**, 697 (2025).
50. Díez Menéndez, C., Poczai, P., Williams, B., Myllys, L. & Amiryousefi, A. IRplus: An Augmented Tool to Detect Inverted Repeats in Plastid Genomes. *Genome Biol. Evol.* **15**, evad177 (2023).

51. Krämer, C. et al. Removal of the large inverted repeat from the plastid genome reveals gene dosage effects and leads to increased genome copy number. *Nat. Plants*. **10**, 923–935 (2024).
52. Turudić, A. et al. Variation in Chloroplast Genome Size: Biological Phenomena and Technological Artifacts. *Plants* **12**, 254 (2023).
53. Gurdon, C. & Maliga, P. Two Distinct Plastid Genome Configurations and Unprecedented Intraspecies Length Variation in the *accD* Coding Region in *Medicago truncatula*. *DNA Res. Int. J. Rapid Publ Rep. Genes Genomes*. **21**, 417–427 (2014).
54. Liu, T. et al. Complete Chloroplast Genome Sequence Structure and Phylogenetic Analysis of *Brassica juncea* var. *multiceps* (Brassicaceae). *Agronomy* **15**, (2025).
55. Martiansyah, I. et al. Characterization of the complete chloroplast genome of *Hopea rudiformis* P.S. Ashton, a critically endangered and endemic tree from Borneo Island. *Sci. Technol.* **0**, 1–13 (2025).
56. Kaila, T. et al. Chloroplast Genome Sequence of Clusterbean (*Cyamopsis tetragonoloba* L.): Genome Structure and Comparative Analysis. *Genes* **8**, 212 (2017).
57. Adhikari, B., Parajuli, S. & Nepal, M. P. Reporting complete chloroplast genome of endangered red mulberry, useful for understanding hybridization and phylogenetic relationships. *Sci. Rep.* **15**, 13403 (2025).
58. Sadhu, L. et al. Chloroplasts evolved an additional layer of translational regulation based on non-AUG start codons for proteins with different turnover rates. *Sci. Rep.* **13**, 896 (2023).
59. Zanduetta-Criado, A. & Bock, R. Surprising features of plastid *ndhD* transcripts: addition of non-encoded nucleotides and polysome association of mRNAs with an unedited start codon. *Nucleic Acids Res.* **32**, 542–550 (2004).
60. Barthet, M. M., Pierpont, C. L. & Tavernier, E. Unraveling the role of the enigmatic MatK maturase in chloroplast group IIA intron excision. *Plant. Direct.* **4**, e00208 (2020).
61. Hildebrand, M., Hallick, R. B., Passavant, C. W. & Bourque, D. P. Trans-splicing in chloroplasts: the *rps* 12 loci of *Nicotiana tabacum*. *Proc. Natl. Acad. Sci.* **85**, 372–376 (1988).
62. 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**, 194 (2021).
63. Holwerda, B. C., Jana, S. & Crosby, W. L. CHLOROPLAST AND MITOCHONDRIAL DNA VARIATION IN *HORDEUM VULGARE* AND *HORDEUM SPONTANEUM*. *Genetics* **114**, 1271–1291 (1986).
64. Vanichanon, A., Blake, N., Sherman, J. & Talbert, L. Multiple origins of allopolyploid *Aegilops triuncialis*. *Theor. Appl. Genet.* **106**, 804–810 (2003).
65. Powell, W. et al. Hypervariable microsatellites provide a general source of polymorphic DNA markers for the chloroplast genome. *Curr. Biol. CB.* **5**, 1023–1029 (1995).
66. Cauz-Santos, L. A. Beyond conservation: the landscape of chloroplast genome rearrangements in angiosperms. *New. Phytol.* **247**, 2571–2580 (2025).
67. Sabir, J. et al. Evolutionary and biotechnology implications of plastid genome variation in the inverted-repeat-lacking clade of legumes. *Plant. Biotechnol. J.* **12**, 743–754 (2014).

68. Doyle, J. J., Doyle, J. L., Ballenger, J. A. & Palmer, J. D. The distribution and phylogenetic significance of a 50-kb chloroplast DNA inversion in the flowering plant family Leguminosae. *Mol. Phylogenet Evol.* **5**, 429–438 (1996).
69. Stein, D. B., Palmer, J. D. & Thompson, W. F. Structural evolution and flip-flop recombination of chloroplast DNA in the fern genus *Osmunda*. *Curr. Genet.* **10**, 835–841 (1986).
70. Jansen, R. K., Wojciechowski, M. F., Sanniyasi, E., Lee, S. B. & Daniell, H. Complete plastid genome sequence of the chickpea (*Cicer arietinum*) and the phylogenetic distribution of *rps12* and *clpP* intron losses among legumes (Leguminosae). *Mol. Phylogenet Evol.* **48**, 1204–1217 (2008).
71. Chen, Y. et al. Comparative chloroplast genomics of Rutaceae: structural divergence, adaptive evolution, and phylogenomic implications. *Front Plant. Sci* **16**, (2025).
72. Moghaddam, M., Ohta, A., Shimizu, M., Terauchi, R. & Kazempour-Osaloo, S. The complete chloroplast genome of *Onobrychis gaubae* (Fabaceae-Papilionoideae): comparative analysis with related IR-lacking clade species. *BMC Plant. Biol.* **22**, 75 (2022).
73. Dong, W. et al. *ycf1*, the most promising plastid DNA barcode of land plants. *Sci. Rep.* **5**, 8348 (2015).
74. Zhang, Z., Zhang, D. S., Zou, L. & Yao, C. Y. Comparison of chloroplast genomes and phylogenomics in the *Ficus sarmentosa* complex (Moraceae). *PloS One.* **17**, e0279849 (2022).
75. Univaso, L., Peña, F., Román-Figueroa, C. & Paneque, M. The *rps15-rpl32* intergenic locus is a phylogeographic marker for Latin American *Zea mays* landrace varieties and subspecies. *BMC Genom.* **26**, 672 (2025).
76. Magee, A. M. et al. Localized hypermutation and associated gene losses in legume chloroplast genomes. *Genome Res.* **20**, 1700–1710 (2010).
77. Shaw, J. et al. The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *Am. J. Bot.* **92**, 142–166 (2005).
78. Taberlet, P., Gielly, L., Pautou, G. & Bouvet, J. Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant. Mol. Biol.* **17**, 1105–1109 (1991).
79. Liu, X. et al. Complete Chloroplast Genome Sequence and Phylogenetic Analysis of *Quercus bawanglingensis* Huang, Li et Xing, a Vulnerable Oak Tree in China. *Forests* **10**, (2019).
80. Andueza-Noh, R. H. et al. Multiple domestications of the Mesoamerican gene pool of lima bean (*Phaseolus lunatus* L.): evidence from chloroplast DNA sequences. *Genet. Resour. Crop Evol.* **60**, 1069–1086 (2013).
81. Mishra, B., Gautam, M., Thakur, N. & Chahota, R. K. Phylogenetic Relationship Among Cultivated and Wild Species of Genus *Macrotyloma* Based on Cytology and Molecular Markers. *Plant. Mol. Biol. Rep.* **43**, 458–473 (2025).
82. Parks, M., Cronn, R. & Liston, A. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biol.* **7**, 84 (2009).

83. Oyeibanji, O., Zhang, R., Chen, S. Y. & Yi, T. S. New Insights Into the Plastome Evolution of the Millettoid/Phaseoloid Clade (Papilionoideae, Leguminosae). *Front Plant. Sci* **11**, (2020).
84. Wojciechowski, M. F., Lavin, M. & Sanderson, M. J. A phylogeny of legumes (Leguminosae) based on analysis of the plastid matK gene resolves many well-supported subclades within the family. *Am. J. Bot.* **91**, 1846–1862 (2004).
85. Mahesh, H. B. et al. Genome, Transcriptome, and Germplasm Sequencing Uncovers Functional Variation in the Warm-Season Grain Legume Horsegram *Macrotyloma uniflorum* (Lam.) Verdc. *Front Plant. Sci* **12**, (2021).
86. Cheung, T. Y., Kafoutchoni, K. M., Agoyi, E. E., Chan, T. F. & Chapman, M. A. Exceptionally low genomic diversity in the underutilised legume Kersting's groundnut. *Nat. Commun.* **16**, 5183 (2025).
87. Egan, A. N., Vatanparast, M. & Cagle, W. Parsing polyphyletic *Pueraria*: Delimiting distinct evolutionary lineages through phylogeny. *Mol. Phylogenet Evol.* **104**, 44–59 (2016).

Figures

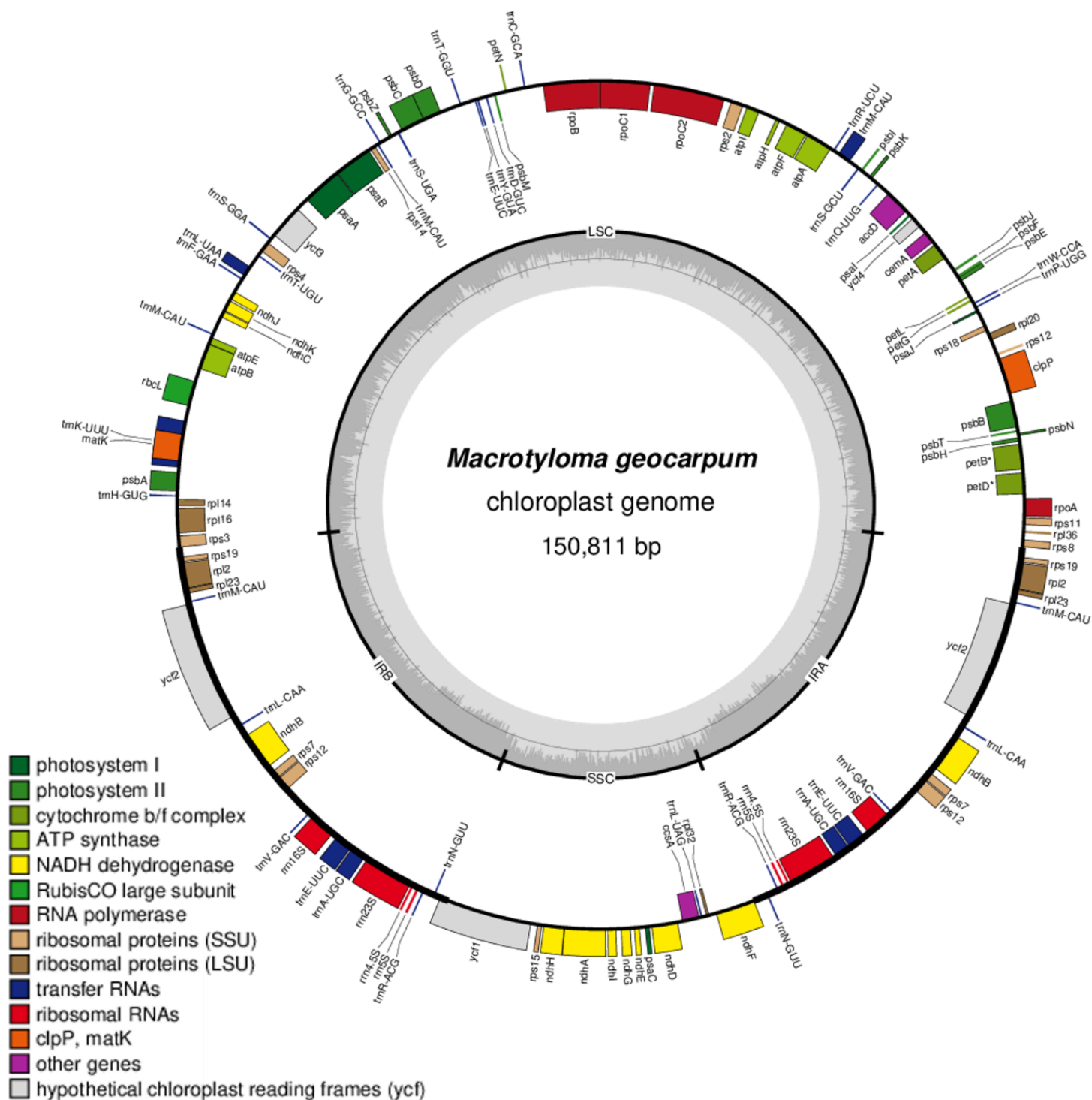


Figure 1

Circular map of the chloroplast genome of *Macrotyloma geocarpum*

The circular map illustrates the complete chloroplast genome structure of the *Macrotyloma geocarpum*, showing the typical quadripartite organisation comprising a large single-copy (LSC) region, a small single-copy (SSC) region, and two inverted repeat regions (IRA and IRB). Genes located inside the circle are transcribed clockwise, while those on the outside are transcribed counterclockwise. Genes are colour-coded by functional category, including photosystem genes, ribosomal proteins (LSU and SSU),

transfer RNAs, ribosomal RNAs, *clpP/matK*, hypothetical chloroplast ORFs, and other functional genes. The inner grey histogram represents the variation in GC content across the genome. The boundaries of the LSC, SSC, and IR regions are indicated.

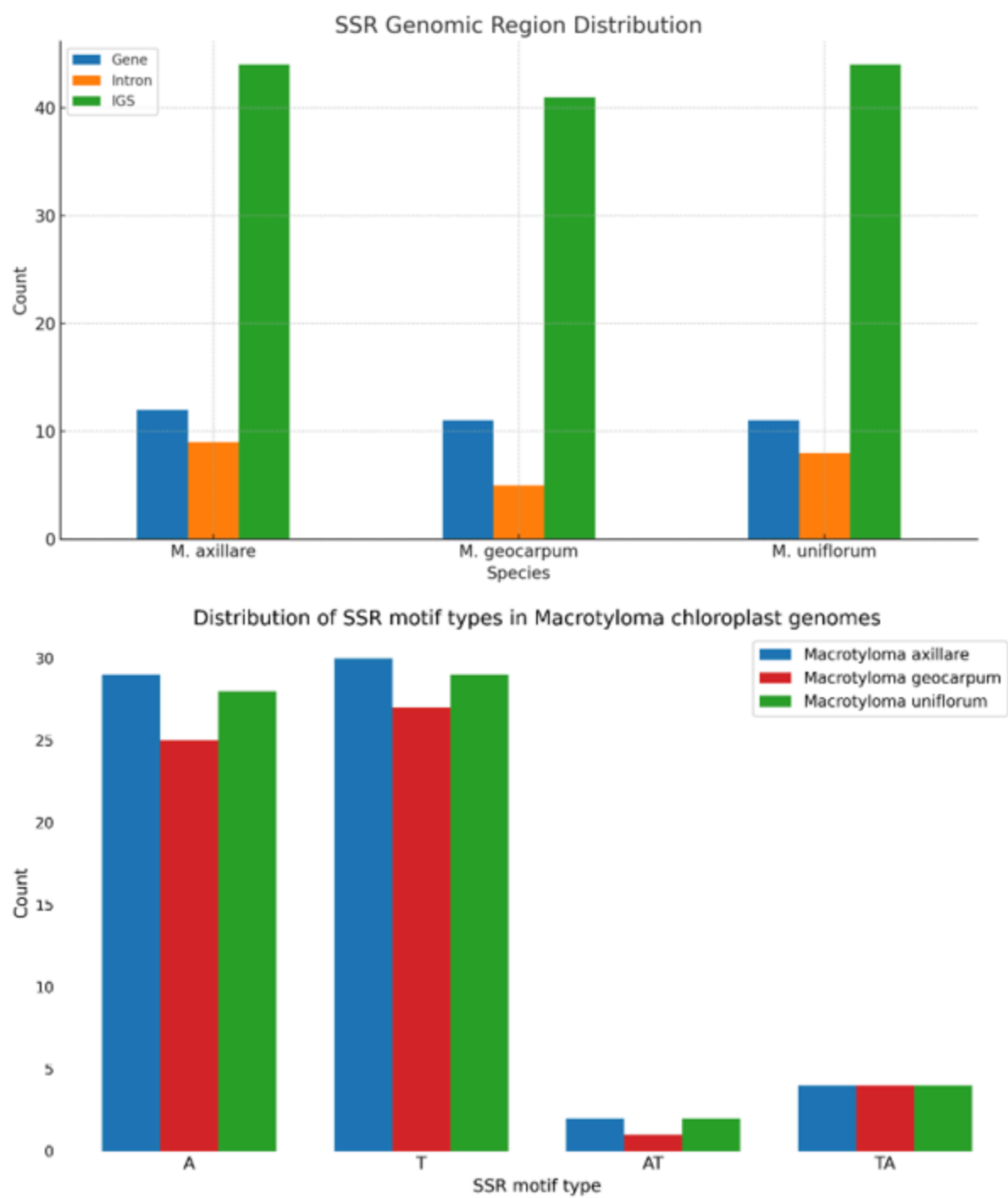


Figure 2

A Distribution of simple sequence repeats (SSRs) across genomic regions of three *Macrotyloma* chloroplast genomes. The bar chart shows the count of SSRs located within gene regions, introns, and

intergenic spacers (IGS) of *M. axillare*, *M. geocarpum*, and *M. uniflorum*. In all three species, the majority of SSRs occur in the IGS regions, followed by the gene regions and introns. This pattern suggests that SSRs are predominantly located in non-coding regions of the chloroplast genome.

B. Frequency and motif type distribution of SSRs detected in the chloroplast genomes of three *Macrotyloma* species. The figure presents the number of SSRs belonging to major motif types (A, AT, T, TA, and others) across *M. axillare*, *M. geocarpum*, and *M. uniflorum*. Mononucleotide A/T repeats are the most abundant category in all three species, followed by dinucleotide repeats (primarily TA). Minor motif types appear at low frequencies.

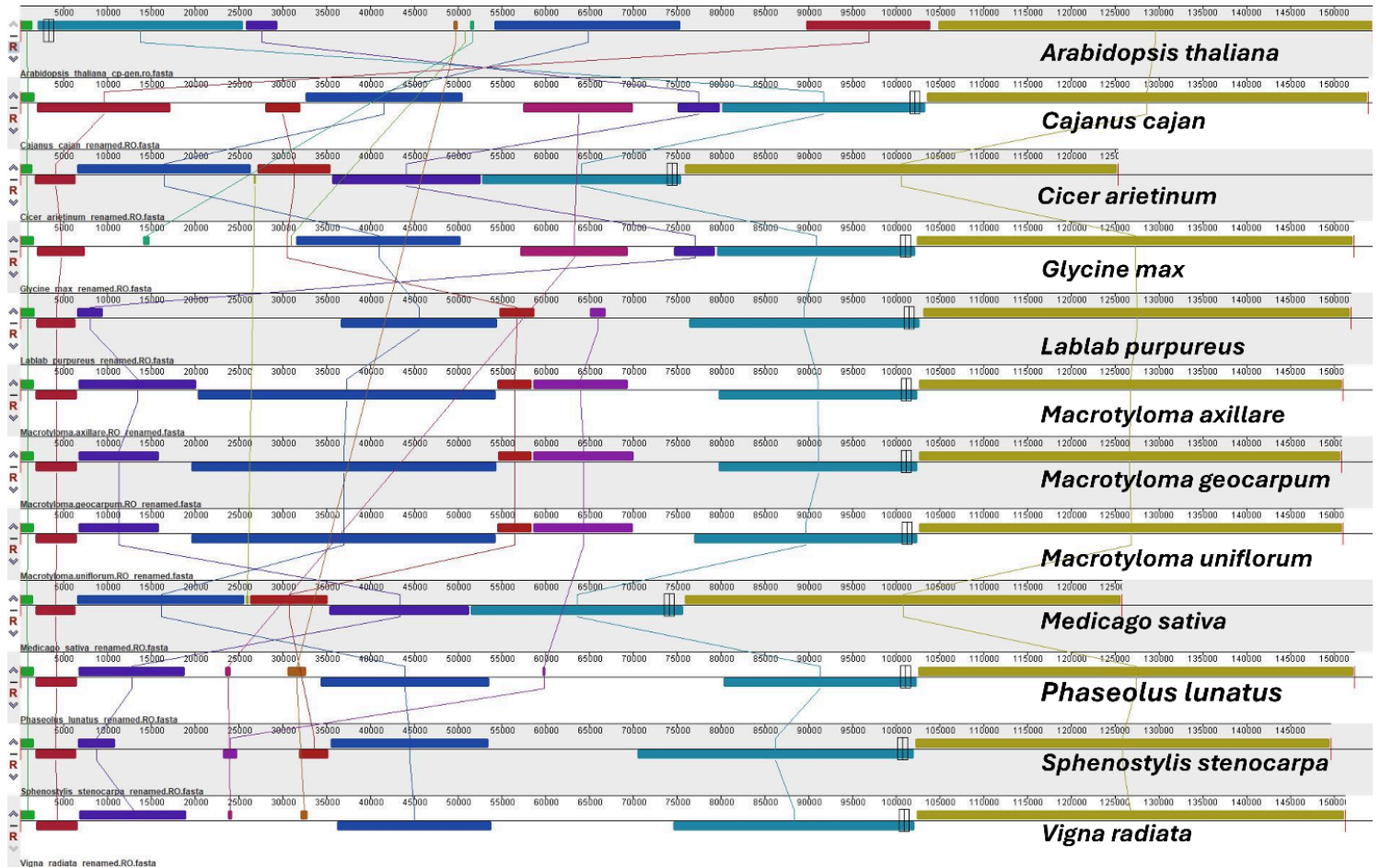


Figure 3

Whole-chloroplast genome alignment of seven chloroplast genomes, including three *Macrotyloma* species, using the Mauve to identify locally collinear blocks (LCBs). Coloured blocks represent syntenic regions shared among the genomes, with each block appearing in a unique colour. The height of the coloured regions indicates the level of sequence similarity within each LCB. Blocks positioned above or below the centre line indicate forward or reverse orientation, respectively, revealing conserved regions and inversion events. Black connecting lines between genomes highlight rearrangements, shifts in gene order, and boundaries of homologous regions. Gene annotations are shown as open boxes, enabling visualisation of gene-level collinearity and structural divergence.



Figure 4

Comparison of chloroplast genome structure and junction boundaries among three *Macrotyloma* species and representative members of the Phaseoleae tribe. The figure illustrates the quadripartite chloroplast genome organisation, including the Large Single-Copy (LSC), Small Single-Copy (SSC), and the two Inverted Repeat regions (IRa and IRb), across eight Phaseoleae species. The positions and lengths of genomic regions are shown along with the distances between key genes and the four junctions: JLB (LSC-IRb), JSB (IRb-SSC), JSA (SSC-IRa), and JLA (IRa-LSC).

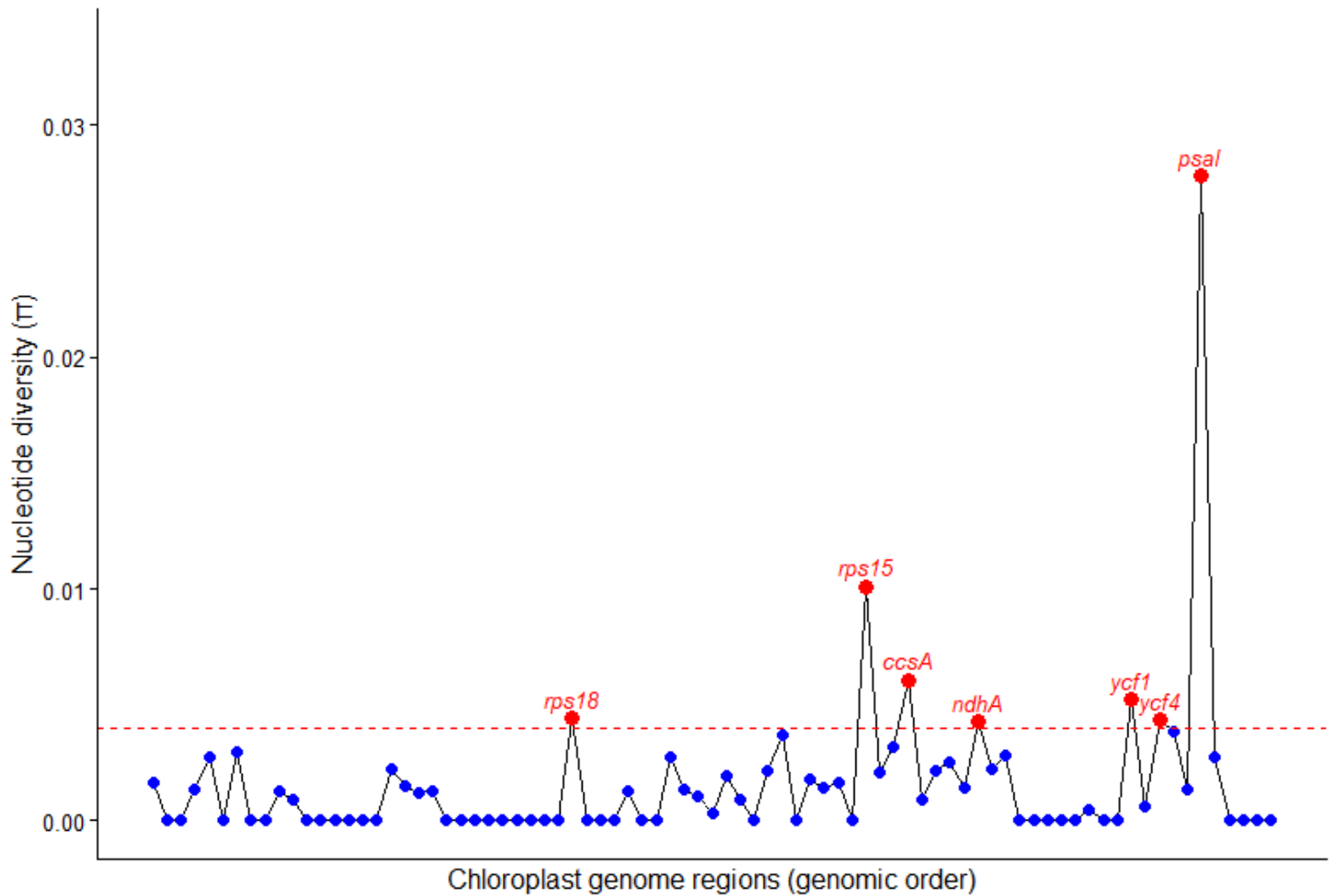


Figure 5

Gene-wise nucleotide diversity (π) across the chloroplast genome of *Macrotyloma* species. Blue points represent π values for each gene, while the black line connects adjacent genes to illustrate variation patterns across the plastome. Genes showing elevated nucleotide diversity relative to the genome-wide mean are highlighted in red and annotated, including *rps18*, *rps15*, *ccsA*, *ndhA*, *ycf4*, *ycf1*, and *psal*.

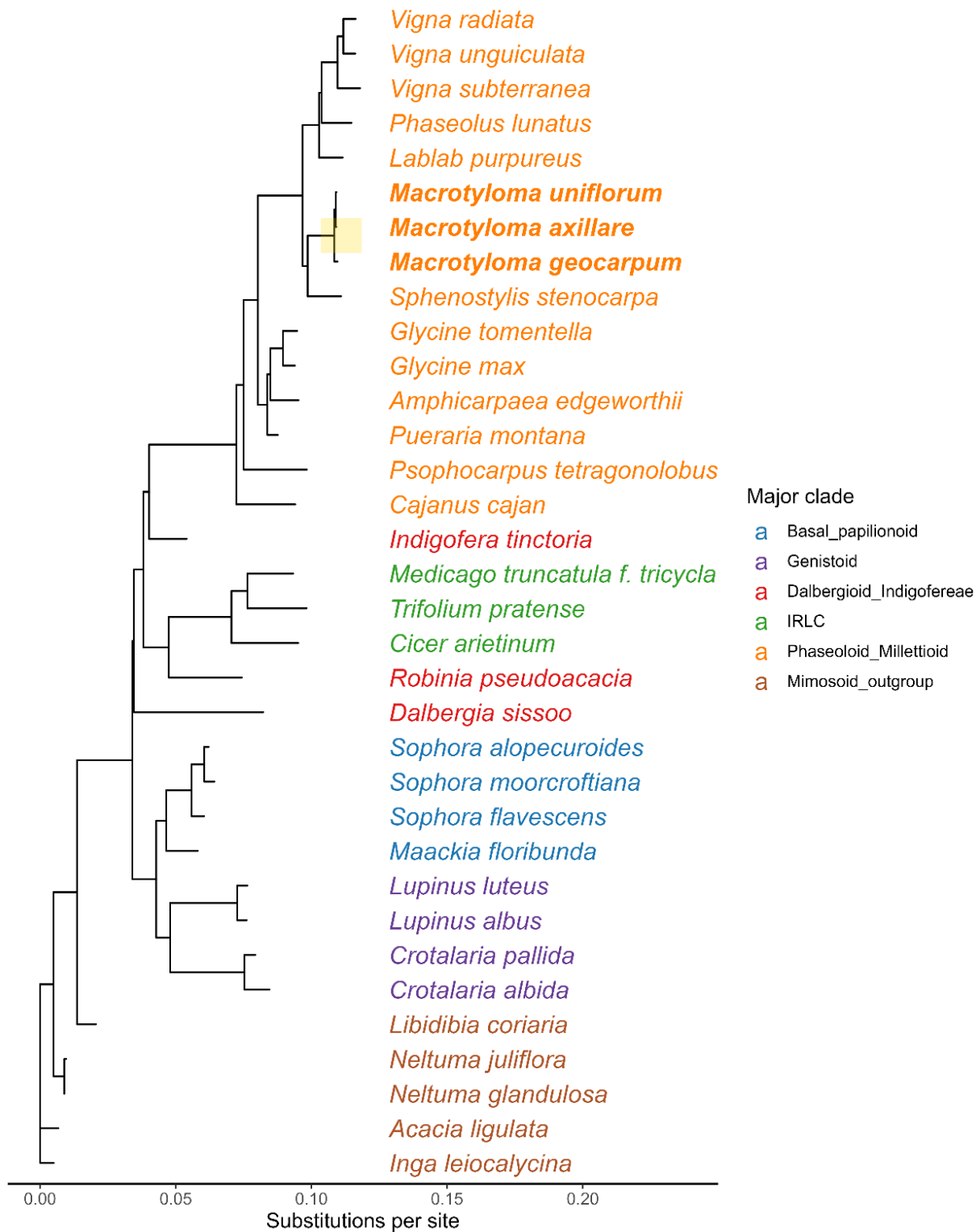


Figure 6

Maximum-likelihood phylogeny of 42 shared chloroplast protein-coding genes from 34 species representing major lineages of the Fabaceae family. The tree was reconstructed using concatenated CDS sequences under a maximum-likelihood framework. Branch lengths are proportional to the number of substitutions per site (scale bar = 0.01). Species names are colour-coded by major tribal or subfamily groups for visual clarity. The tree is rooted with Caesalpinioideae outgroups.

Supplementary Files

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

- [Supplementarytables.xlsx](#)
- [Supplementaryfigure.docx](#)