

RESEARCH

Open Access



Comparative genomics and phylogenetic analysis of six Malvaceae species based on chloroplast genomes

Yiwang Zhong^{1,2†}, Beibei Bai^{1†}, Yangyang Sun^{1,2}, Ke Wen^{1,3}, Yang Qiao^{2,4}, Lijun Guo^{2,3}, Huidong Deng^{2,3}, Yingjun Ye^{2,4}, Liying Feng^{5*} and Xuejie Feng^{1,2*}

Abstract

The Malvaceae family, comprising 9 subfamilies and 4,225 species, includes economically significant taxa, such as *Ceiba pentandra*, *Gossypium ekmanianum*, *Gossypium stephensii*, *Kokia drynarioides*, *Talipariti hamabo*, and *Durio zibethinus*. Chloroplast (cp) genome research is crucial for elucidating the evolutionary divergence and species identification within this family. In this study, we assembled and annotated cp genomes of six Malvaceae species, conducting comprehensive comparative genomic and phylogenomic analyses. The assembled genomes range from 160,495 to 163,970 bp in size, with 125–129 genes annotated. Notable differences were observed in the IR (inverted repeat) regions, and SSR analysis revealed that *Durio zibethinus* has the highest number of specific variation sites. Among the six species, *Talipariti hamabo* uniquely exhibits more palindromic repeats than forward repeats. Seven highly mutated regions were identified, offering potential markers for species identification. Phylogenetic reconstruction using the maximum likelihood method revealed two primary clades within Malvaceae: Byttneriina and Malvadendrina. Within Malvadendrina, the subfamily Helicteroideae represents the earliest divergence, followed by Sterculioideae. This study provides a robust phylogenetic framework and valuable insights into the classification and evolutionary history of Malvaceae species.

Keywords Chloroplast genome, Malvaceae, Sequence characteristic, Comparative analysis, Phylogeny

[†]Yiwang Zhong and Beibei Bai contributed equally to this work.

*Correspondence:

Liying Feng

lyfengad@163.com

Xuejie Feng

13807680898@163.com

¹Sanya Institute, Hainan Academy of Agricultural Sciences, Sanya 572024, China

²Institute of Tropical Fruit Trees, Hainan Academy of Agricultural Sciences, Haikou 571100, China

³Key Laboratory of Genetic Resources Evaluation and Utilization of Tropical Fruits and Vegetables (Co-construction by Ministry and Province), Ministry of Agriculture and Rural Affairs, Haikou 571100, China

⁴Key Laboratory of Tropical Fruit Tree Biology of Hainan Province, Haikou 571100, China

⁵Yazhouwan National Laboratory, Sanya 572024, China

Introduction

Chloroplasts (cp), vital organelles for photosynthesis, are ubiquitously found in plants and eukaryotic algae [1, 2]. With the advent of the genomics era, significant progress has been made in the study of cp genomes [3]. Despite its relatively compact size, the cp genome demonstrates a remarkable balance between conservation and variability in both gene content and structural organization across diverse plant species [3, 4]. The cp genome typically exhibits a quadripartite structure, comprising a large single-copy (LSC) region, a small single-copy (SSC) region, and two inverted repeat (IR) regions. This structure encodes between 110 and 130 genes, with genome sizes ranging from 110 to 240 kb [5, 6]. Compared to the



mitochondrial and nuclear genomes, the cp genome is generally inherited and does not undergo recombination [7]. Although the structure of the cp genome is generally conserved in angiosperms, significant variations have been observed in genome size, structural organization, and gene substitution rates [8, 9].

Owing to its mode of maternal inheritance, the cp genome offers direct insights into the evolutionary history of maternal lineages in plants [10, 11]. Furthermore, the conserved gene sequences and structural characteristics of the cp genome make it an ideal molecular marker for investigating large-scale phylogenetic relationships among plants [12, 13]. The evolution of the cp genome also reflects the history of plant adaptation to environmental changes [14]. The loss, rearrangement, and horizontal transfer of certain genes reveal the mechanisms by which plants have adapted to different ecological environments throughout their evolutionary history [15, 16]. This information is pivotal for elucidating plant diversity, evolutionary mechanisms, and ecological adaptability. Consequently, the investigation of the cp genome not only deepens our understanding of photosynthesis and cellular functions in plants but also supplies critical molecular evidence for studies on plant evolution, biodiversity, and ecological adaptation [17, 18].

Malvaceae family, also known as the mallow family, comprises approximately 244 genera and around 4,225 species, including well-known plants such as okra, cotton, cacao, roses, and durian [19–21]. The species of the Malvaceae family are widely distributed across tropical, subtropical, and temperate regions of the world. They are found in diverse habitats, ranging from forests and grasslands to arid and coastal areas [22]. The Malvaceae family is distinguished by its extensive variety of species, ranging from diminutive herbaceous plants to towering trees. And they are of great ornamental, economic, edible and medicinal importance. Such as the flowers, typically exhibiting five petals and vibrant hues, are a hallmark of the family and are frequently cultivated in gardens and landscape design [23–25]. Additionally, such as cotton, which is a major source of natural fiber [26], Durian is renowned as the “King of Fruits” due to its distinctive aroma and creamy texture [27]. *Ceiba pentandra*, native to tropical America, is widely cultivated across tropical regions worldwide. It is extensively distributed in the tropical areas of Asia and Africa, with particularly abundant populations in Indonesia [28]. Some species of the Malvaceae family have been used in traditional medicine for centuries. Known for its anti-inflammatory, antioxidant, and liver-protective properties, *Malva sylvestris* is used to treat respiratory diseases, skin inflammation, and gastrointestinal problems [29]. The fibers from Velvetleaf (*Abutilon theophrasti*) possess excellent strength and durability, and they are commonly used for making ropes

and fabrics [30]. In molecular biology research, plants of the Malvaceae family, due to their diversity and wide distribution, have become important models for studying plant evolution, genomics, and ecology [23]. With advancements in genomics and bioinformatics technologies, research on the phylogeny, gene functions, and metabolic pathways of Malvaceae plants will continue to deepen, enhancing our understanding and application of plant biology.

In the past few decades, phylogenetic research on the Malvaceae family has made significant progress [21, 23]. Traditional morphological classification methods have gradually been supplemented and replaced by molecular systematics studies, which primarily utilize sequence data from nuclear genes, cp genes, and mitochondrial genes [31–33]. The Malvaceae family is divided into nine subfamilies: Byttnerioideae, Grewioideae, Helicteroideae, Sterculioideae, Brownlowioideae, Dombeyoideae, Tilioideae, Bombacoideae, and Malvoideae [19]. Studies based on DNA fragments have found that Bombacoideae and Malvoideae form a sister clade, while Byttnerioideae and Grewioideae form a sister clade with the remaining subfamilies [34–37]. Studies based on cp genomes suggest that Byttnerioideae and Grewioideae form a distinct clade, while the remaining seven subfamilies form another clade [38–40]. Currently, the datasets cover relatively few species within each subfamily. Therefore, expanding the species-level sampling of Malvaceae and conducting phylogenetic studies based on cp genomes can better clarify the evolutionary relationships within the family.

In this study, we reported the cp genomes of six Malvaceae species, six species have had their nuclear genomes published, with species identification already confirmed in these publications [25, 41–44]. Although chloroplast genomes of several durian cultivars have been published [45–47], the complete chloroplast genome data for the ‘Monthong’ cultivar remain unclear. Similarly, the chloroplast genome data for the other five species also lack consistency or completeness. A comparative analysis was conducted across these six species, elucidating structural variations in each. Additionally, leveraging cp genome data, we identified the hotspot regions of variation in the cp genome and elucidated the phylogenetic relationships among the nine subfamilies of the Malvaceae. This study will provide valuable insights for clarifying species identification and the phylogenetic history of Malvaceae plants.

Materials and methods

Plant material sampling and DNA extraction

The data for the five species used in this study were downloaded from National Center for Biotechnology Information (<https://www.ncbi.nlm.nih.gov/>) on 10 January 2024. One species was sequenced from our own sample

collection. The data for *Ceiba pentandra* (PRJNA905654) was obtained from the Institute of Cotton Research Yacheng Base, Chinese Academy of Agricultural Sciences, Sanya, Hainan Province, China (18°23′43.3896″N, 109°10′27.4620″E). Samples of *Gossypium ekmanianum* and *Gossypium stephensii* (PRJNA739494) were sourced from the National Wild Cotton Nursery in Sanya, China, managed by the Institute of Cotton Research, Chinese Academy of Agricultural Sciences (ICR-CAAS). Fresh leaf tissue from *Kokia drynarioides* (PRJNA400144) was collected from the Iowa State University greenhouse. *Talipariti hamabo* (PRJNA759075) tissue was obtained from the Institute of Botany, Jiangsu Province, Chinese Academy of Sciences. The *Durio zibethinus* samples were collected from the Yucai durian plantation in Sanya, Hainan. For *D. zibethinus*, young, developing leaf tissues from healthy plants were immediately frozen in liquid nitrogen and stored at -80 °C to preserve DNA integrity. Total genomic DNA was extracted using the CTAB method [48]. The genome was sequenced on the Illumina NovaSeq 6000 platform using next-generation sequencing technology, with paired-end reads of 150 bp and an insert size of 350 bp. Genome indexing and alignment were performed using Bowtie2 V2.4.1 [49] and SAMtools V1.13 [50], coverage and depth calculations were conducted using the bamdst V1.1.0 (<https://github.com/shiquan/bamdst>).

Characterization of cp genomes in six Malvaceae species

The sequencing data were subjected to quality control and filtering using Fastp V0.15.0 [51]. After filtering and screening the sequencing data, the cp genome was automatically assembled using GetOrganelle V1.7.1 [8] with k-mer parameters set to 21, 45, 75, 85, and 105. The assembly was visualized using Bandage V0.8.1 [52], which facilitated the removal of redundant contigs and the circularization of the sequence. The assembled cp genome was annotated using CPGAVAS2 [53], and a circular map was generated using OGDRAW [54].

Simple repeat sequence annotation

Simple sequence repeats (SSRs), including mononucleotide, dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide SSRs, were predicted using MISA [55]. The parameters were set to >10 repeat units for mononucleotide SSRs, >6 repeat units for dinucleotide SSRs, ≥5 repeat units for trinucleotide SSRs, 24 repeat units for tetranucleotide SSRs, and ≥3 repeat units for both pentanucleotide and hexanucleotide SSRs.

Characterization and comparative analysis of genomes

The GC content of each cp genome, including the LSC, SSC, and IR regions, was determined using Extractseq V6.6.0 and Seqkit V2.3.1. The number of genes in the

total cp genome, tRNA, and rRNA was determined from the GB files of each VP genome. Repetitive sequences, including forward (F), reverse (R), complement (C), and palindromic (P) repeats, were identified using REPuter [56] with a minimum repeat size set to 30 bp and a Hamming distance of 3. Gene loss was visualized using the heatmap function from the R package. The cp genome sequences were compared using the mVISTA program [57]. IRscope V0.1 [58] was used to visualize the contraction and expansion of the IR boundaries between the four regions of the genome (LSC/IRb/SSC/IRa). Sequence alignment was performed using MAFFT V7.310 [59], and a sliding window analysis was conducted with DnaSP V6.12.03 [60] to calculate Pi across the cp genome. The step size was set to 200 bp, with a window length of 600 bp.

Phylogenetic analysis

To elucidate the phylogenetic relationships within the Malvaceae family, 141 cp genome sequences of Malvaceae were downloaded from the NCBI database (Table S2). Using *Dipterocarpus turbinatus* (NC_046842.1) as an outgroup, a phylogenetic analysis was conducted on a total of 148 cp genomes, including six species assembled in this study. We first corrected the SSC regions of all cp genomes by reversing them. After aligning the sequences consistently for all species, we constructed the phylogenetic tree. Multiple sequence alignments were performed using MAFFT V7.310 [59], were trimmed using TrimAl V1.4 [61] with a minimum conservation threshold and a gap threshold of 20%. Phylogenetic reconstruction was conducted using the Maximum Likelihood (ML) method. The ML tree was generated with 1000 bootstrap replicates using the GTRGAMMA model, implemented in RAxML V8.1.17 [62]. The analyses were run with the command “raxmlHPC-PTHREADS-SSE3 -f a -x 12345 -p 12345 -s all.fasta -m GTRGAMMA -N 1000 -n output”. We downloaded the protein files of the nuclear genomes for 22 species and identified single-copy orthologous genes and constructed species tree using OrthoFinder V3.0 [63]. iTOL (<https://itol.embl.de/>) was used to enhance and annotate the evolutionary tree, as well as to incorporate subgroup information.

Result

Characterization of cp genomes in six Malvaceae species

In this study, we assembled the cp genomes of six species within the Malvaceae family. These cp genomes exhibit the typical quadripartite structure, comprising a LSC region, a SSC region, and two IR regions (Fig. 1). The genome sizes range from 160,495 to 163,970 bp, GC content of 35.83–37.34%, and the number of genes identified varies between 125 and 129 (Table 1). The GC content of the IR regions is significantly higher than that of the LSC

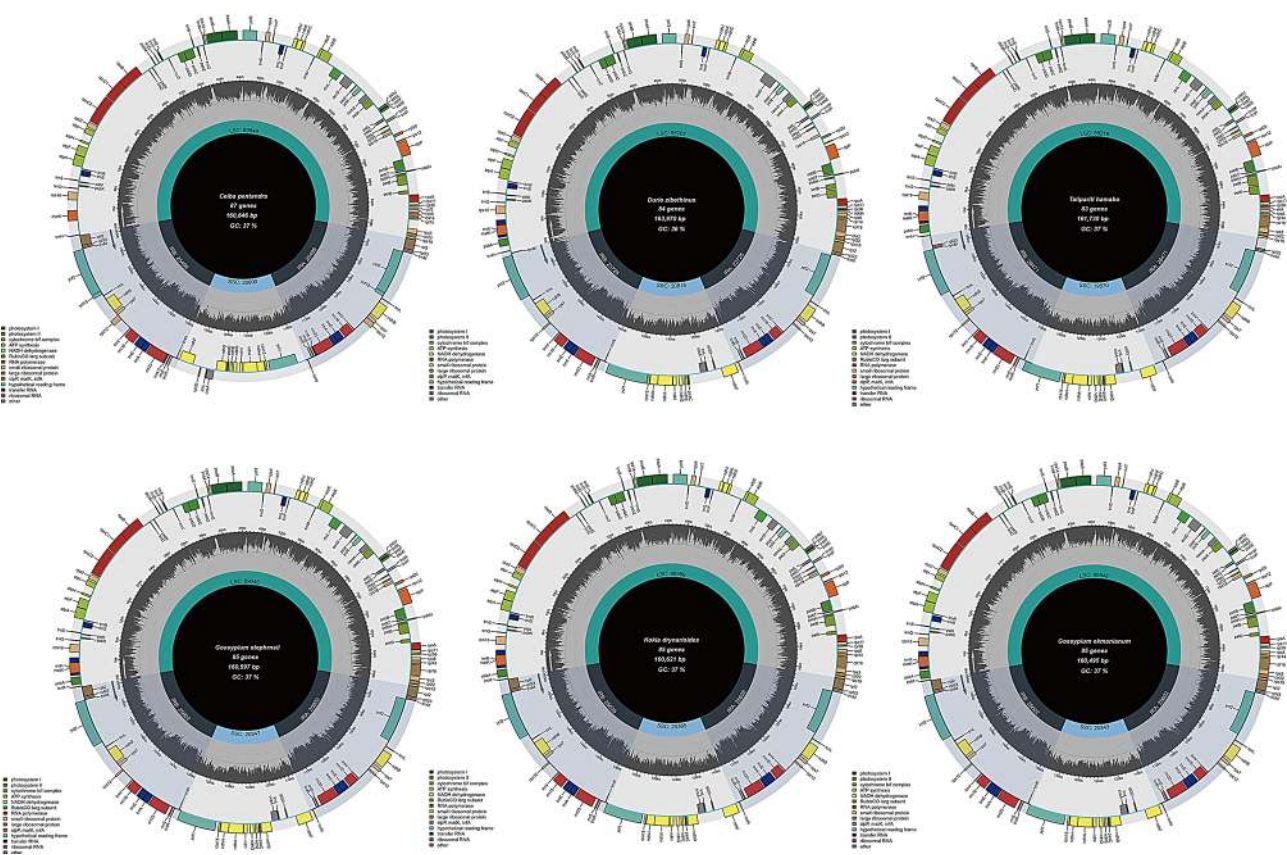


Fig. 1 Cp genome maps of six Malvaceae. Genes shown outside of the outer layer circle are transcribed clockwise, while those inside are transcribed counterclockwise. The genes belonging to different functional groups are color-coded. The dark gray area of the inner circle denotes

Table 1 Comparison of cp features among six Malvaceae plants

	<i>C. pentandra</i>	<i>D. zibethinus</i>	<i>G. ekmanianum</i>	<i>G. stephensii</i>	<i>K. drynarioides</i>	<i>T. hamabo</i>
Total length (bp)	160,646	163,970	160,495	160,597	160,621	161,730
GC (%)	36.81	35.83	37.19	37.18	37.34	36.9
LSC.length (bp)	89,648	95,702	88,948	89,046	88,986	89,218
SSC.length (bp)	20,000	20,818	20,343	20,347	20,365	19,570
IR.length (bp)	50,998	47,450	51,204	51,204	51,270	52,942
Total number of genes	129	128	127	127	127	125
Protein-coding genes	87	84	85	85	85	83
tRNA genes	36	36	36	36	36	36
rRNA genes	6	8	6	6	6	6

and SSC regions. This pattern of GC content distribution is consistent with observations in other plant species [14, 64]. Among the six species, *T. hamabo* has the fewest protein-coding genes, with a total of 83. *D. zibethinus* contains the highest amount of rRNA with 8. The content of tRNA (36) is consistent across all species.

Codon bias and sequence repeats analysis

The patterns of codon usage and nucleotide composition contribute to establishing a theoretical foundation for the genetic modification of cp genomes [65]. The codon usage frequency of protein-coding genes in

the cp genomes of six assembled Malvaceae species was thoroughly analyzed. These genomes contain between 25,128 and 26,256 codons, with *D. zibethinus* having the fewest and *G. ekmanianum* and *G. stephensii* containing the most. Among all the codons, leucine (Leu) emerged as the most prevalent amino acid, with a frequency ranging from 10.25 to 10.40%. This was followed by isoleucine (Ile), with a frequency between 10.39% and 10.62%. Conversely, cysteine (Cys) was the least abundant, with a frequency of 1.13–1.15% (Fig. 2A, Table S3). After logarithmically transforming the usage counts of all amino acids, statistical analysis and clustering were performed.

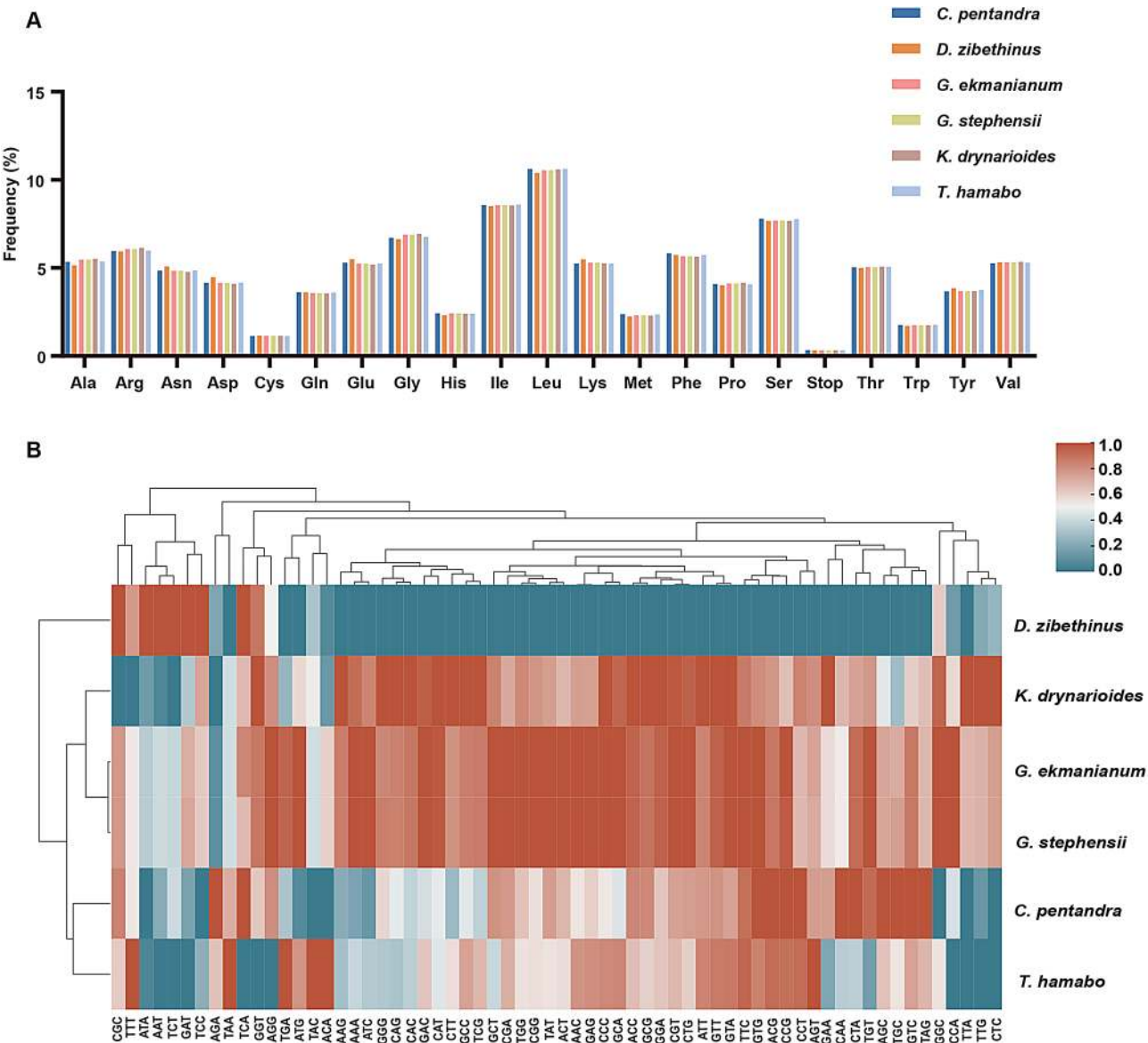


Fig. 2 Analysis of codon preferences in the cp genomes of six Malvaceae species. **A.** amino acid usage; **B.** codon usage

The results revealed differences in the usage of different amino acids encoded by the same codon. *D. zibethinus* exhibited significant differences compared to the other five species. Furthermore, the stop codon TAA was preferentially used (Fig. 2B, Table S4).

We further conducted an analysis of SSR and repetitive sequences in six species. Upon statistical analysis of single nucleotide polymorphisms, we found that *C. pentandra* exhibited the highest number, while *K. drynarioides* had the lowest. Notably, except for *C. pentandra*, tetra-nucleotides were absent in the other five species. The overall SSR analysis revealed that *D. zibethinus* contained the highest number of SSRs (96) (Fig. 3A). Repetitive sequence analysis indicated that *D. zibethinus* had a significantly higher number of forward repeats compared to

the other five species, whereas *K. drynarioides* exhibited the highest number of complement repeats (Fig. 3B).

Comparative analysis between six Malvaceae species

The expansion and contraction of the IR region are major driving forces in the evolution of land plants [66–69]. In this study, we compared the adjacent genes and boundary conditions of the four regions (LSC-IRb-SSC-IRa) among six species (Fig. 4). The results revealed that the *rps19* gene is located at the LSC and IRb boundary, but this gene is absent in durian compared to the other five species. The *ndhF* gene is situated in the SSC region, with a significant reduction in the nucleotide sequence of the *ndhF* gene in *T. hamabo*. The *ycf1* gene spans the SSC and IRa boundary, with 963 bp of this gene located in the IRa

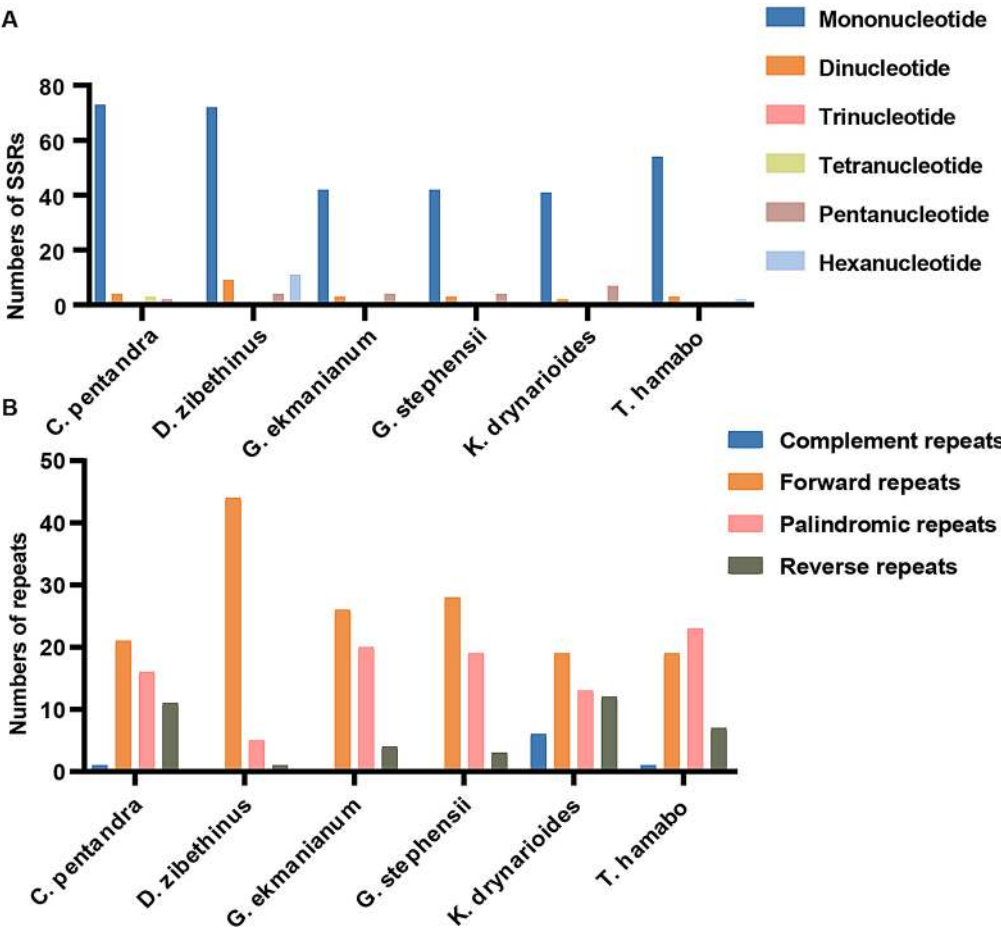


Fig. 3 Analysis of sequence repeat in the six Malvaceae cp genomes. **A.** The type frequency of different SSR types. **B.** Frequency of identified long repeats types

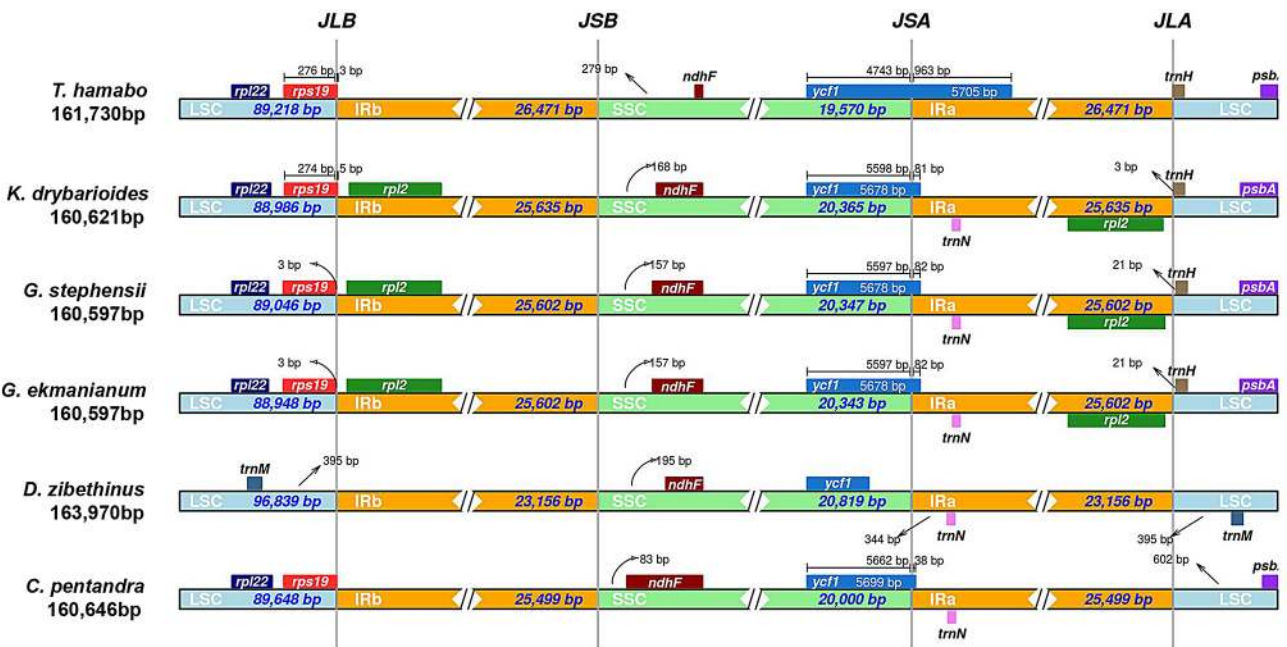


Fig. 4 Comparison of the borders of the LSC, SSC, and IR regions among six Malvaceae cp genomes

region in *T. hamabo*, whereas in *D. zibethinus*, the entire gene is situated in the SSC region. In the other four species, the *ycf1* gene extends less than 100 bp into the IRA region. The genome size expansion in *D. zibethinus* is primarily observed in the LSC region. The *rpl2* gene is only present in *K. drynarioides*, *G. stephensii*, and *G. ekmanianum*. Additionally, the inverted repeat regions in *D. zibethinus* are the shortest compared to the other five species.

Through phylogenetic analysis and gene loss statistics of six species, we found that *C. pentandra* diverged first, followed by *D. zibethinus*, then *T. hamabo*. The two species of the genus *Gossypium* form sister branches (Fig. 5). The gene loss statistics revealed that the gene with the highest frequency in the six species is *trnM-CAU* (4 copies), followed by *trnY-GUA* (3 copies). Additionally, we found that *D. zibethinus* exhibits gene loss in *clpP* and *rpl32*, and *rpl2* is lost in *T. hamabo* and *C. pentandra*. There is only one copy of *rpl2* in *D. zibethinus*, whereas the remaining three species have two copies. Both *D. zibethinus* and *C. pentandra* have two *ycf15* genes, while the other four species show a loss of this gene (Fig. 5; Table S5).

We also analyzed the differences in the cp genomes of six species and further explored their genetic relationships (Fig. 6). Using mVISTA for global sequence alignment, with *D. zibethinus* sequence and annotation file as references, we observed variations across different regions among the six species. The results indicated that the cp genomes in Malvaceae species are highly conserved, with few variable regions, consistent with previous studies [66–68]. Additionally, we found that the IR regions are more conserved compared to the LSC and SSC regions, exhibiting fewer variable sites. Most variations were observed in the conserved non-coding sequences (NCS), with coding regions showing less divergence than non-coding regions. Among the coding genomes, genes with significant sequence differences included *rpl2*, *ycf2*, *ndhF*, and *ndhD*.

Identification of variability hotspots

We performed whole cp genome analyses on six newly assembled species using DnaSP V6.12.03 [60], calculating Pi and Theta values (Fig. 7; Table S6). The mean Pi value

of the aligned gene sequences was 0.03708, and the mean Theta value was 0.04431. We identified seven highly variable regions: *petA-pabJ*, *rbcL-accD*, *rpl32-trnL-UAG*, *rps16-trnQ-UUG*, *psbC-trnS-UGA*, *ndhF-rpl32*, and *ycf1*. Of these, four were located in the LSC region, while the remaining three were in the SSC region (Fig. 7). This pattern aligns with previous findings indicating that the IR region is typically more conserved than the LSC and SSC regions.

Phylogenetic reconstruction of Malvaceae

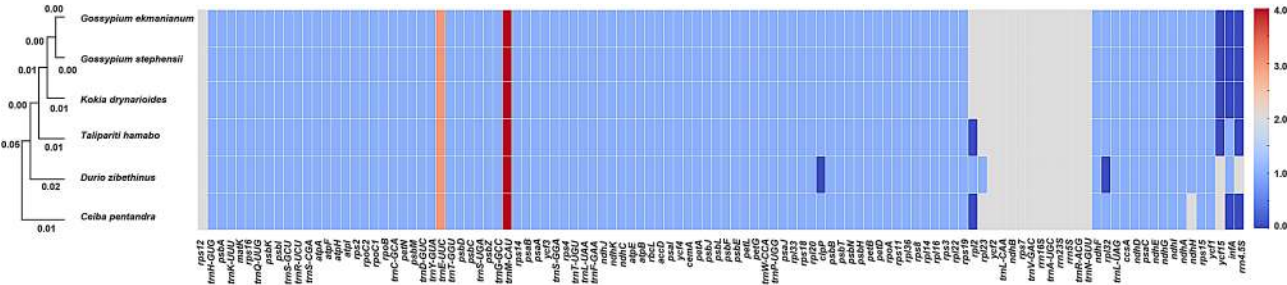
To elucidate the phylogenetic relationships within the Malvaceae family, we constructed a phylogenetic tree using the Maximum Likelihood (ML) method based on 148 genomes, including all genes and intergenic spacers (Fig. 8; Fig. S2). The results show that Byttnerioideae and Grewioideae form one clade, while the remaining seven subfamilies cluster into another clade. The phylogenetic relationships among the subfamilies Brownlowioideae, Tilioideae, and Dombeyoideae are closer. *K. drynarioides* is more closely related to *Thespesia populnea*. Genus *Talipariti* clusters together with *Hibiscus* and *Urena*, while *C. pentandra* forms a clade with the genus *Pachira*. *G. ekmanianum* and *G. stephensii* form a small clade. The topology of the phylogenetic tree has strong support at each node. All branches had support rates above 70%.

To validate the accuracy of the phylogenetic tree, we selected the nuclear genomes of 24 Malvaceae species and one outgroup species that are currently available online. Using homologous genes, we reconstructed a species tree (Fig. 9; Table S7). Grewioideae diverged first as the basal clade, followed by Byttnerioideae.

Discussion

Chloroplast genome features in Malvaceae

The comparative analysis of plant genomes offers novel insights into patterns of molecular evolution cp sustain life on Earth by converting solar energy into carbohydrates through photosynthesis and oxygen release [70–72]. The advent of next-generation sequencing (NGS) technologies has equipped scientists with more rapid and cost-efficient methods for sequencing cp genomes [73, 74]. These advancements have significantly accelerated



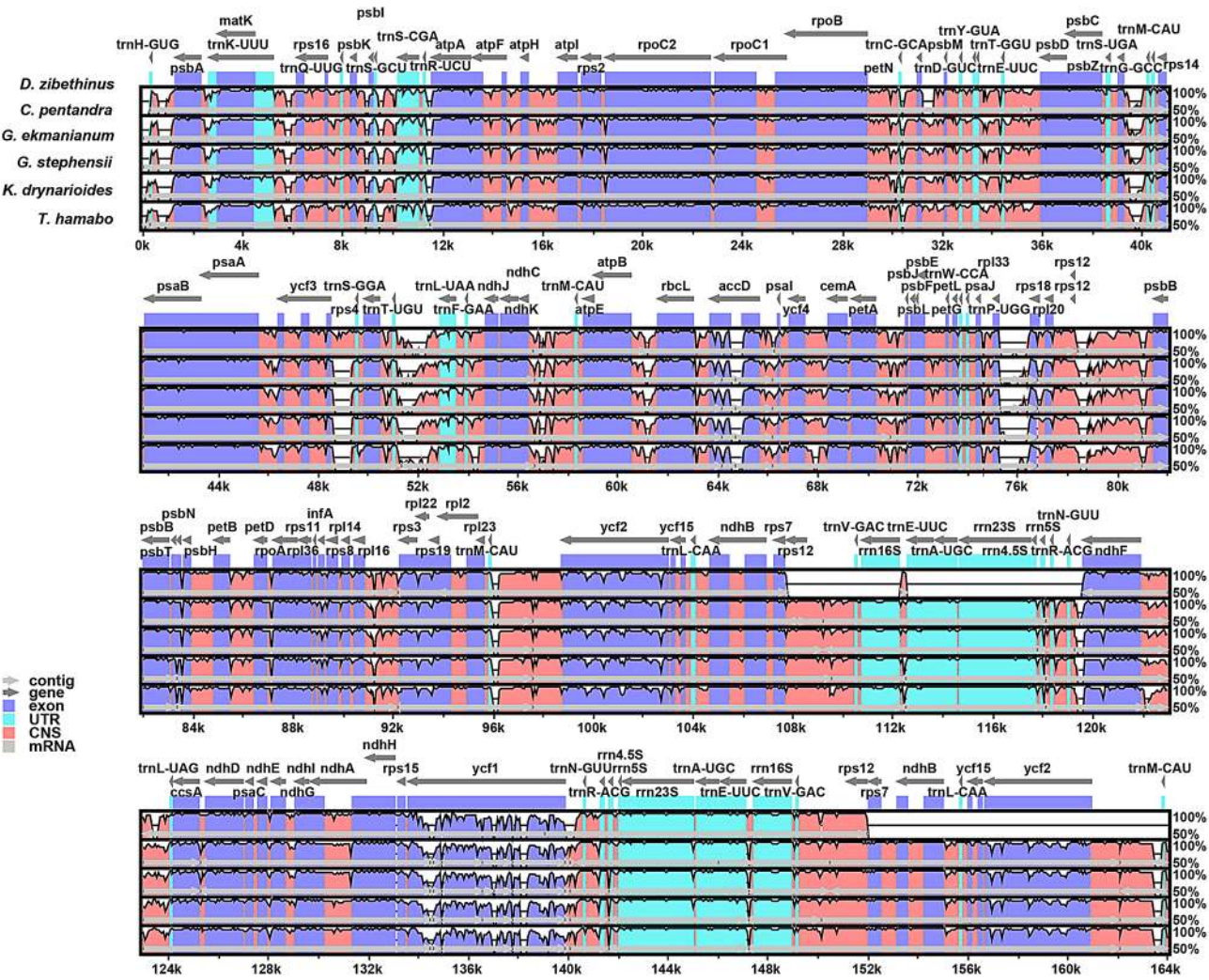


Fig. 6 Analysis of sequence alignment of six Malvaceae cp genomes

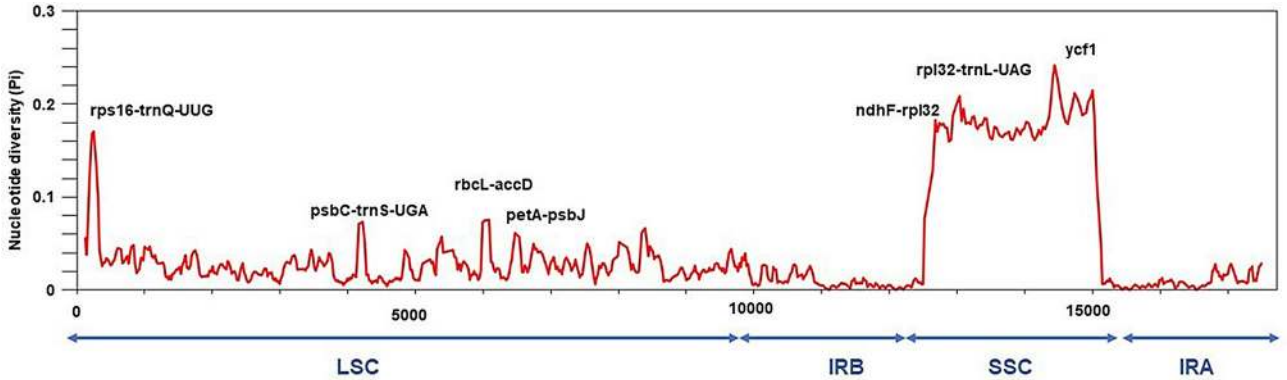


Fig. 7 Nucleotide polymorphism of the cp genomes of six Malvaceae species

genomic research, enabling more comprehensive and detailed analyses of cp genome. This study utilized NGS data to obtain a detailed picture of the cp genome structures of six species within the Malvaceae. The GC content varies across the LSC, IR, and SSC regions, with the

IR region having the highest GC content. The higher GC content in the IR region may be attributed to the abundance of rRNA genes in this region [75, 76]. Angiosperms exhibit a conserved quadripartite structure comprising two IR regions, a LSC region, and a SSC region, with

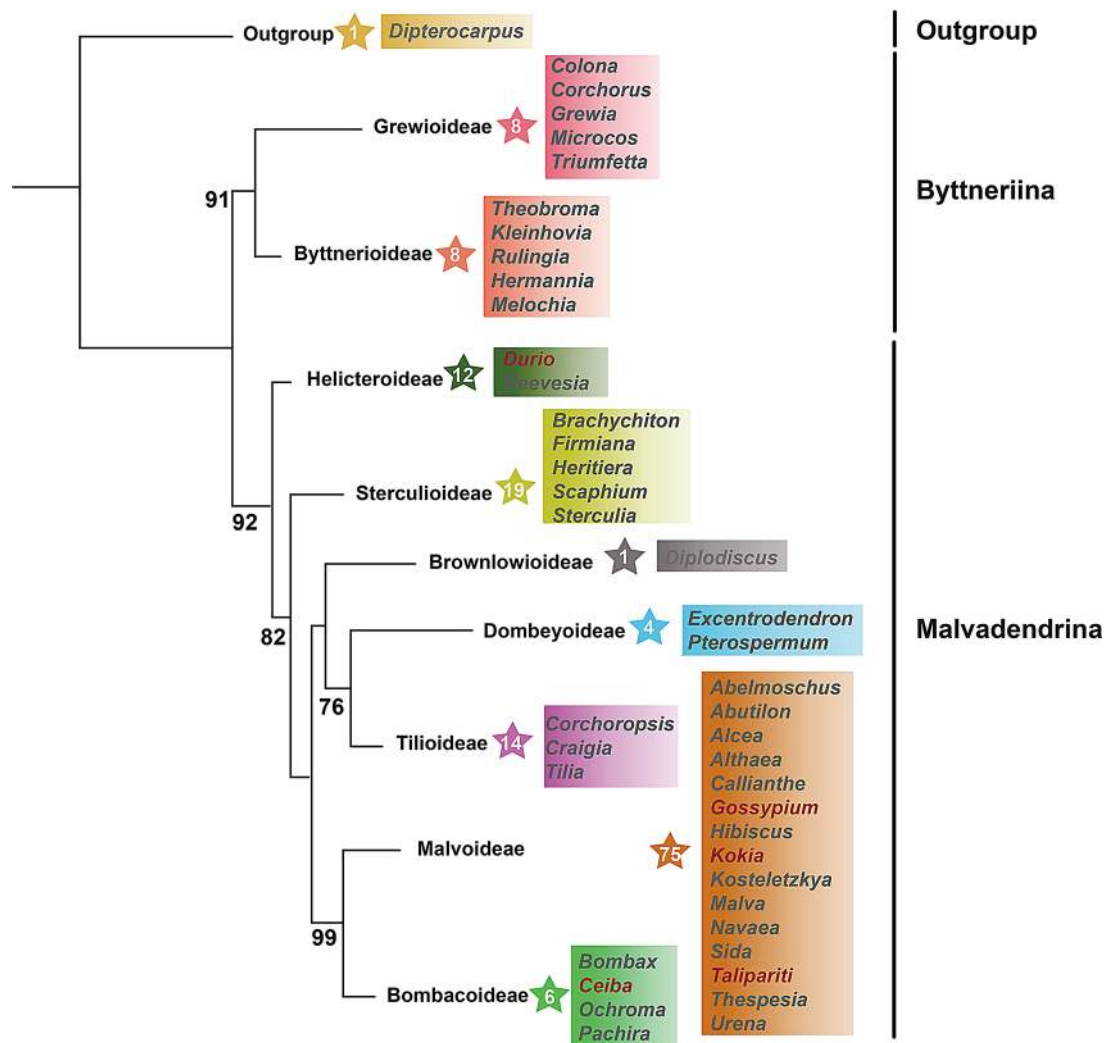


Fig. 8 Phylogenetic relationships among subfamilies with complete cp sequence of Malvaceae. Species in genus assembled in this study are indicated in red text. The numbers within star symbols represent the number of species included

genome sizes ranging from 110 kb to 240 kb [77–79]. The cp genomes of the six Malvaceae species we assembled range from 160,495 to 163,970 bp (Fig. 1), GC content ranges between 35.83% and 37.34%, indicating the highly conserved nature of cp. genomes in Malvaceae species.

In the structure of cp genomes, the IR region is typically the most conserved. Due to the highly conserved nature of cp. genomes, gene loss within this organelle provides crucial insights into species evolution [80]. Gene loss primarily occurs in *rpl2* and *ycf1* genes. In this study, cp genome exhibited a loss of the *rpl32* gene in *D. zibethinus* (Fig. 4). This phenomenon may be attributed to abnormal DNA replication, repair, or recombination during the evolutionary process of their common ancestor [81, 82].

Potential molecular markers in Malvaceae

Although the cp. genome evolves slowly and is relatively conserved, it still contains several mutational hotspots frequently used for species identification and molecular markers, such as *rbcL*, *ndhF*, *matK*, and *psbA-trnH* [76, 83]. These loci provide valuable genetic variation for phylogenetic studies and biodiversity assessments. The IR region is crucial for stabilizing the structure of the cp genome. Compared to the two single-copy regions, its slower nucleotide substitution rate enhances the efficacy of copy correction mechanisms [84]. Through the assembly of six cp. genomes, we found that the variability in the SSC and LSC regions is higher than that in the IR regions, and the variability in non-coding regions is higher than that in coding regions [11, 85]. These findings are consistent with studies of cp genomes in other higher plants. Additionally, we identified seven highly variable regions, among which the *ycf1* gene shows potential for

Tree scale 0.02

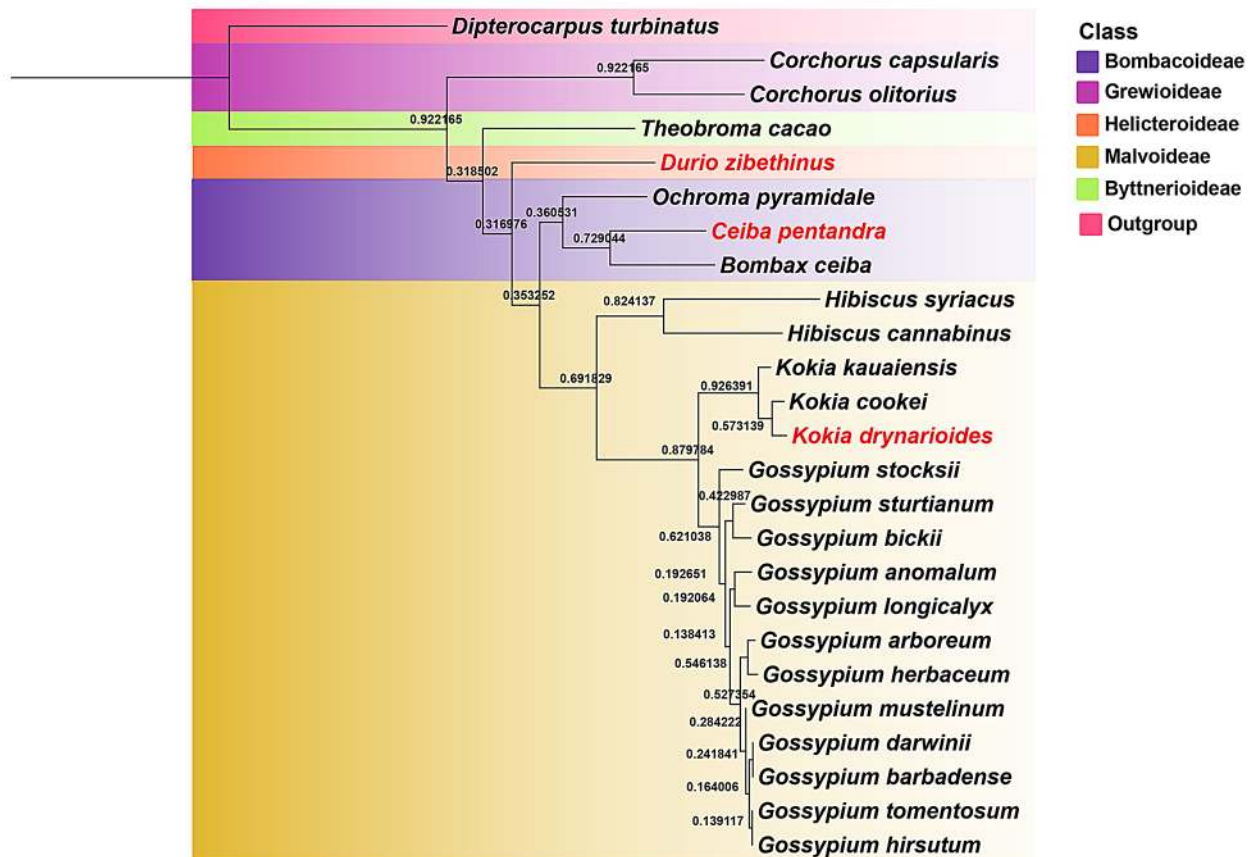


Fig. 9 The phylogenetic relationships constructed based on homologous genes derived from the protein sequences of 25 nuclear genomes

species identification. These results also indicate that the non-coding regions in the Malvaceae evolve at a faster rate than the coding regions. Therefore, when selecting molecular markers for species in the Malvaceae, non-coding regions should be prioritized.

Repetitive sequences can enhance genetic diversity within species and influence cpDNA rearrangements [40, 86]. Gene duplication serves as a significant source of organelle evolution and novel genetic functions [87]. Duplications in the single-copy regions are typically triggered by expansions of the IR regions. Overall, cpSSRs are valuable genetic tools for population genetics and evolutionary studies due to their codominance, high polymorphism, and low substitution rates [88, 89]. In our study, we identified a total of 50–96 SSRs across the six species (Fig. 3). SSRs are valuable genetic markers for distinguishing closely related species. In the six species we studied, mononucleotide SSRs were the most prevalent, aligning with observations in other Malvaceae species [4, 40]. These SSRs also hold potential for the development of lineage-specific cpSSR markers [90].

Phylogenetic relationships in Malvaceae

Cp genome contains essential genes and is often used in biotechnology or phylogenetic research [17]. In the Malvadendrina clade, the phylogenetic relationship between Malvoideae and Bombacoideae was the first to be established [34]. Currently, the phylogenetic relationships within the Malvaceae family remain contentious. Most phylogenetic studies on Malvaceae suggest that Helicteroideae occupies a basal position within the Malvadendrina clade. Phylogenies based on single genes or a few genes often suffer from limited sample sizes, resulting in topological inconsistencies and low support values [35]. Cp genome data confirmed a close relationship among Brownlowioideae, Tilioideae, and Dombeyoideae, with Brownlowioideae represented by a single sequence. The analysis further demonstrated that Brownlowioideae is the sister group to the other two subfamilies [23]. In this study, we employed complete cp genome sequences to perform a comprehensive phylogenetic analysis of the relationships among nine subfamilies within the Malvaceae family. This analysis has refined and updated the phylogenetic positioning of the six Malvaceae species included in our research.

The entire Malvaceae family is primarily divided into two major clades, Byttneriina and Malvanderina, consistent with previously reported phylogenetic relationships [24, 37, 91]. In the phylogenetic studies of the Malvaceae family, most research indicates that the subfamily Helicteroideae is situated at the base of the Malvanderina clade. However, analyses based on individual cp. genes suggest that Dombeyoideae is the earliest-diverging basal group. Our study supports the positioning of the subfamily Helicteroideae at the base of the Malvanderina clade (Fig. 8).

Conclusion

This study employed Illumina sequencing data to assemble and annotate the cp genomes of six species within the Malvaceae family, offering novel insights into the evolutionary patterns of their cp genomes. Compared to other Malvaceae species, the cp genomes of these taxa demonstrate considerable conservation in terms of genome size, structural organization, and gene content, with the notable exception of a highly variable *ycf1* gene. The molecular data presented herein provide an invaluable resource for advancing the understanding of evolutionary processes within the Malvaceae family.

Abbreviations

CP	Chloroplast
DNASP	DNA sequence polymorphism
IR	Inverted repeat
LSC	Large single copy
ML	Maximum likelihood
PI	Nucleotide diversity
SSC	Small single copy
SSR	Simple sequence repeat

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12870-024-05974-w>.

Supplementary Material 1
Supplementary Material 2
Supplementary Material 3
Supplementary Material 4
Supplementary Material 5
Supplementary Material 6
Supplementary Material 7
Supplementary Material 8
Supplementary Material 9
Supplementary Material 10

Acknowledgements

The authors would like to thank Yingjun Ye from Institute of Tropical Fruit Trees, Hainan Academy of Agricultural Sciences for revising the manuscript.

Author contributions

XJF, LYF and YWZ designed the research study. YWZ and BBB contributed materials. YWZ and BBB annotated and analyzed the genomes, wrote the manuscript. YYS and YQ isolated DNA, assembled the genomes. LJG, KW, HDD and LYF helped YWZ to analyze the data. The final manuscript was approved by all authors.

Funding

This work was supported by Hainan Academy of Agricultural Sciences program (HAAS2022JBGS01), Key Project of the Tropical Fruit Research Institute, Hainan Academy of Agricultural Sciences (2024SJD01), Hainan Provincial Key Research and Development Project (ZDYF2024XDNY251), Research Initiation Fund Project for Recruited Talents at the Tropical Fruit Research Institute, Hainan Academy of Agricultural Sciences (ITFT2024RCOD03), and Sanya science and technology innovation special project (2022KJCX11).

Data availability

The study of new sequencing data has been deposited in the NCBI (<https://www.ncbi.nlm.nih.gov/>), accession number is PRJNA1130351.

Declarations

Ethical approval

The study was conducted the plant material that complies with relevant institutional, national, and international guidelines and legislation.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 30 October 2024 / Accepted: 16 December 2024

Published online: 26 December 2024

References

- Keeling PJ. The endosymbiotic origin, diversification and fate of plastids. *Philos Trans R Soc Lond B Biol Sci.* 2010;365(1541):729–48.
- Green BR. Chloroplast genomes of photosynthetic eukaryotes. *Plant J.* 2011;66(1):34–44.
- Daniell H, Lin CS, Yu M, Chang WJ. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biol.* 2016;17(1):134.
- Xu Y, Fang B, Li J, Wang Y, Liu J, Liu C, et al. Phylogenomic analysis and development of molecular markers for the determination of twelve plum cultivars (*Prunus*, Rosaceae). *BMC Genomics.* 2022;23(1):745.
- Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ, Boore JL, et al. The complete chloroplast genome sequence of *Pelargonium hortorum*: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. *Mol Biol Evol.* 2006;23(11):2175–90.
- Jiang D, Cai X, Gong M, Xia M, Xing H, Dong S, et al. Complete chloroplast genomes provide insights into evolution and phylogeny of *Zingiber* (Zingiberaceae). *BMC Genomics.* 2023;24(1):30.
- Hu G, Wu Y, Guo C, Lu D, Dong N, Chen B, et al. Haplotype analysis of Chloroplast genomes for Jujube breeding. *Front Plant Sci.* 2022;13:841767.
- Jin JJ, Yu WB, Yang JB, Song Y, DePamphilis CW, Yi TS, et al. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* 2020;21(1):241.
- Ruhlman TA, Jansen RK. The plastid genomes of flowering plants. *Maliga P, editor Chloroplast Biotechnology: methods and protocols.* Totowa, NJ: Humana; 2014. pp. 3–38.
- Kim Y, Shin J, Oh D, Kim A, Choi C. Comparative Analysis of Complete Chloroplast Genome Sequences and insertion-deletion (Indel) Polymorphisms to Distinguish five *Vaccinium* species. *Forests.* 2020;11(9):927.
- Liu H, Ye H, Zhang N, Ma J, Wang J, Hu G, et al. Comparative analyses of chloroplast genomes provide Comprehensive insights into the adaptive evolution of *Paphiopedilum* (Orchidaceae). *Horticulturae.* 2022;8:391.
- Palmer JD. Comparative organization of chloroplast genomes. *Annu Rev Genet.* 1985;19:325–54.

13. Olmstead R, Palmer J. Chloroplast DNA systematics: a review of methods and data analysis. *Am J Bot*. 1994;81:1205–24.
14. Zhang Y, Wang Z, Guo Y, Chen S, Xu X, Wang R. Complete chloroplast genomes of *Leptodermis Scabrida* complex: comparative genomic analyses and phylogenetic relationships. *Gene*. 2021;791:145715.
15. Gao C, Li T, Zhao X, Wu C, Zhang Q, Zhao X, et al. Comparative analysis of the chloroplast genomes of *Rosa* species and RNA editing analysis. *BMC Plant Biol*. 2023;23(1):318.
16. Liu F, Chen N, Wang H, Li J, Wang J, Qu F. Novel insights into chloroplast genome evolution in the green macroalgal genus *Ulva* (Ulvophyceae, Chlorophyta). *Front Plant Sci*. 2023;14:1126175.
17. Zhang D, Ren J, Jiang H, Wang VO, Dong X, Hu G. Comparative and phylogenetic analysis of the complete chloroplast genomes of six *Polygonatum* species (Asparagaceae). *Sci Rep*. 2023;13(1):7237.
18. Gu X, Li L, Zhong X, Su Y, Wang T. The size diversity of the Pteridaceae family chloroplast genome is caused by overlong intergenic spacers. *BMC Genomics*. 2024;25(1):396.
19. Bayer C, Fay MF, De Bruijn AY Savolainen, Morton CM, Kubitzki K, et al. Support for an expanded family concept of Malvaceae within a circumscribed order Malvales: a combined analysis of plastid *atpB* and *rbcl* DNA sequences. *Bot J Linn Soc*. 1999;129(4):267–303.
20. de Abreu VAC, Moysés Alves R, Silva SR, Ferro JA, Domingues DS, Miranda VFO, et al. Comparative analyses of *Theobroma cacao* and *T. grandiflorum* mitogenomes reveal conserved gene content embedded within complex and plastic structures. *Gene*. 2023;849:146904.
21. Wu M, He L, Ma G, Zhang K, Yang H, Yang X. The complete chloroplast genome of *Diplodiscus Trichospermus* and phylogenetic position of Brownlowioideae within Malvaceae. *BMC Genomics*. 2023;24(1):571.
22. The APG. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot J Linn Soc*. 2016;181(1):1–20.
23. Cvetković T, Arecas-Berazain F, Hinsinger DD, Thomas DC, Wieringa JJ, Ganesan SK et al. Phylogenomics resolves deep subfamilial relationships in Malvaceae s.l. *G3 (Bethesda)*. 2021;11(7).
24. Wang J, Moore MJ, Wang H, Zhu Z, Wang H. Plastome evolution and phylogenetic relationships among Malvaceae subfamilies. *Gene*. 2021;765:145103.
25. Shao L, Jin S, Chen J, Yang G, Fan R, Zhang Z, et al. High-quality genomes of *Bombax ceiba* and *Ceiba pentandra* provide insights into the evolution of Malvaceae species and differences in their natural fiber development. *Plant Commun*. 2024;5(5):100832.
26. Wang M, Tu L, Yuan D, Zhu D, Shen C, Li J, et al. Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. *Nat Genet*. 2019;51(2):224–9.
27. Li W, Chen X, Yu J, Zhu Y. Upgraded durian genome reveals the role of chromosome reshuffling during ancestral karyotype evolution, lignin biosynthesis regulation, and stress tolerance. *Sci China Life Sci*. 2024;67(6):1266–79.
28. DICK CW, BERMINGHAM E, LEMES MR. Extreme long-distance dispersal of the lowland tropical rainforest tree *Ceiba pentandra* L. (Malvaceae) in Africa and the Neotropics. *Mol Ecol*. 2007;16(14):3039–49.
29. Ekalu A, Habila JD. Phytochemistry, pharmacology and medicinal uses of *Cola* (Malvaceae) family: a review. *Med Chem Res*. 2020;29(12):2089–105.
30. Zhang X, Zhang C, Wang W. Properties of *Abutilon theophrasti* fiber-reinforced high-density polyethylene composites. *BioResources*. 2014;9(2):2876–87.
31. Savolainen V, Chase MW, Hoot SB, Morton CM, Soltis DE, Bayer C, et al. Phylogenetics of flowering plants based on combined analysis of plastid *atpB* and *rbcl* gene sequences. *Syst Biol*. 2000;49(2):306–62.
32. GALTIER N, NABHOLZ B, GLÉMIN S, HURST GDD. Mitochondrial DNA as a marker of molecular diversity: a reappraisal. *Mol Ecol*. 2009;18(22):4541–50.
33. Qian S, Zhang Y, Lee SY. Comparative analysis of complete chloroplast genome sequences in *Edgeworthia* (Thymelaeaceae) and new insights into phylogenetic relationships. *Front Genet*. 2021;12:643552.
34. Baum DA, Dewitt SS, Yen A, Alverson WS, Nyffeler R, Whitlock BA, et al. Phylogenetic relationships of Malvaceae (Bombacoideae and Malvoideae; Malvaceae Ssens lato) as inferred from plastid DNA sequences. *Am J Bot*. 2004;91(11):1863–71.
35. Nyffeler R, Bayer C, Alverson WS, Yen A, Whitlock BA, Chase MW, et al. Phylogenetic analysis of the Malvaceae clade (Malvaceae s.l.) based on plastid DNA sequences. *Organisms Divers Evol*. 2005;5(2):109–23.
36. Won H. Phylogenetic Position of *Corchoropsis* Siebold & Zucc. (Malvaceae s.l.) inferred from Plastid DNA sequences. *J Plant Biology*. 2009;52(5):411–6.
37. Hernández-Gutiérrez R, Magallón S. The timing of Malvales evolution: incorporating its extensive fossil record to inform about lineage diversification. *Mol Phylogenet Evol*. 2019;140:106606.
38. Conover JL, Karimi N, Stenz N, Ané C, Grover CE, Skema C, et al. A Malvaceae mystery: a mallow maelstrom of genome multiplications and maybe misleading methods? *J Integr Plant Biol*. 2019;61(1):12–31.
39. Li J, Ye GY, Liu HL, Wang ZH. Complete chloroplast genomes of three important species, *Abelmoschus moschatus*, *A. manihot* and *A. Sagittifolius*: genome structures, mutational hotspots, comparative and phylogenetic analysis in Malvaceae. *PLoS ONE*. 2020;15(11):e0242591.
40. Li M, Song Y, Sylvester SP, Sylvester SP, Wang X. Comparative analysis of the complete plastid genomes in *Prunus* Subgenus *Cerasus* (Rosaceae): molecular structures and phylogenetic relationships. *PLoS ONE*. 2022;17(4):e0266535.
41. Teh BT, Lim K, Yong CH, Ng CCY, Rao SR, Rajasegaran V, et al. The draft genome of tropical fruit durian (*Durio zibethinus*). *Nat Genet*. 2017;49(11):1633–41.
42. Kayal E, Arick MAI, Hsu C, Thrash A, Yorkston M, Morden CW, et al. Genomic diversity and evolution of the Hawaiian Islands endemic *Kokia* (Malvaceae). *G3*. 2024;14(10):jkae180.
43. Wang Z, Xue JY, Hu SY, Zhang F, Yu R, Chen D, et al. The genome of *Hibiscus hamabo* reveals its adaptation to saline and waterlogged habitat. *Hortic Res*. 2022;9:uhac067.
44. Peng R, Xu Y, Tian S, Unver T, Liu Z, Zhou Z, et al. Evolutionary divergence of duplicated genomes in newly described allotetraploid cottons. *Proc Natl Acad Sci*. 2022;119(39):e2208496119.
45. Shearman JR, Sonthirod C, Naktang C, Sangsrakru D, Yoocha T, Chatbanyong R, et al. Assembly of the durian chloroplast genome using long PacBio reads. *Sci Rep*. 2020;10(1):15980.
46. Cheon SH, Jo S, Kim HW, Kim YK, Sohn JY, Kim KJ. The complete plastome sequence of Durian, *Durio zibethinus* L. (Malvaceae). *Mitochondrial DNA B Resour*. 2017;2(2):763–764.
47. Gia HT, Thi N, Do H, Khang DT. The complete chloroplast genome of *Durio zibethinus* L. Cultivar Ri6 (Helicterioideae, Malvaceae). *Mitochondrial DNA B Resour*. 2024;9(5):625–30.
48. Allen GC, Flores-Vergara MA, Krasynanski S, Kumar S, Thompson WF. A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. *Nat Protoc*. 2006;1(5):2320–5.
49. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9(4):357–9.
50. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–9.
51. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018;34(17):i884–90.
52. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*. 2015;31(20):3350–2.
53. Shi L, Chen H, Jiang M, Wang L, Wu X, Huang L, et al. CPGAVAS2, an integrated plastome sequence annotator and analyzer. *Nucleic Acids Res*. 2019;47(W1):W65–73.
54. Zheng S, Pocai P, Hyvönen J, Tang J, Amiroufesi A. Chloroplot: an online program for the Versatile Plotting of Organelle genomes. *Front Genet*. 2020;11.
55. Thiel T, Michalek W, Varshney R, Graner A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor Appl Genet*. 2003;106(3):411–22.
56. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res*. 2001;29(22):4633–42.
57. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. *Nucleic Acids Res*. 2004;32(suppl2):W273–9.
58. Amiroufesi A, Hyvönen J, Pocai P. Irscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics*. 2018;34(17):3030–1.
59. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;30(4):772–80.
60. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*. 2009;25(11):1451–2.
61. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 2009;25(15):1972–3.
62. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014;30(9):1312–3.

63. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 2019;20(1):238.
64. Parmar R, Cattonaro F, Phillips C, Vassiliev S, Morgante M, Rajora OP. Assembly and Annotation of Red Spruce (*Picea rubens*) Chloroplast Genome, identification of simple sequence repeats, and phylogenetic analysis in *Picea*. *Int J Mol Sci.* 2022.
65. Mazumdar P, Binti Othman R, Mebus K, Ramakrishnan N, Ann Harikrishna J. Codon usage and codon pair patterns in non-grass monocot genomes. *Ann Botany.* 2017;120(6):893–909.
66. Kode V, Mudd EA, Iamtham S, Day A. The tobacco plastid accD gene is essential and is required for leaf development. *Plant J.* 2005;44(2):237–44.
67. Raubeson LA, Peery R, Chumley TW, Dziubek C, Fourcade HM, Boore JL, et al. Comparative chloroplast genomics: analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. *BMC Genomics.* 2007;8:174.
68. Yao X, Tang P, Li Z, Li D, Liu Y, Huang H. The first complete chloroplast genome sequences in Actinidiaceae: genome structure and comparative analysis. *PLoS ONE.* 2015;10(6):e0129347.
69. Abdullah MF, Shahzadi I, Waseem S, Mirza B, Ahmed I, et al. Chloroplast genome of *Hibiscus rosa-sinensis* (Malvaceae): comparative analyses and identification of mutational hotspots. *Genomics.* 2020;112(1):581–91.
70. Roston RL, Jouhet J, Yu F, Gao H. Editorial: structure and function of chloroplasts. *Front Plant Sci.* 2018;9:1656.
71. Hong Z, Wu Z, Zhao K, Yang Z, Zhang N, Guo J et al. Comparative analyses of five complete chloroplast genomes from the Genus *Pterocarpus* (Fabaceae). *Int J Mol Sci.* 2020;21.
72. Zhu XG, Hasanuzzaman M, Jajoo A, Lawson T, Lin R, Liu CM, et al. Improving photosynthesis through multidisciplinary efforts: the next frontier of photosynthesis research. *Front Plant Sci.* 2022;13:967203.
73. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014;30(15):2114–20.
74. Song Y, Feng L, Alyafei M, Jaleel A, Ren M. Function of chloroplasts in plant stress responses. *Int J Mol Sci.* 2021;22(24).
75. Xu Y, Xing M, Li J, Zeng A, Song L, Yan J. Complete chloroplast genome sequence and variation analysis of *Brassica oleracea* L. *Acta Physiol Plant.* 2022;44(10):106.
76. Gao W, Zhou X, Yu Q, Lin G, Fu C, Kang T, et al. Assembly and comparative analyses of the Chloroplast genomes of the threatened plant *Rosa Anemoniflora*. *Forests.* 2024;15:940.
77. Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* 2017;45(4):e18.
78. Zheng S, Hu Y, Zhao R, Yan S, Zhang X, Zhao T, et al. Genome-wide researches and applications on *Dendrobium*. *Planta.* 2018;248(4):769–84.
79. Song Y, Chen Y, Lv J, Xu J, Zhu S, Li M. Comparative chloroplast genomes of *Sorghum* species: sequence divergence and phylogenetic relationships. *Biomed Res Int.* 2019;2019:5046958.
80. Harris ME, Meyer G, Vandergon T, Vandergon VO. Loss of the Acetyl-CoA carboxylase (*accD*) gene in *Poales*. *Plant Mol Biology Report.* 2013;31(1):21–31.
81. Guisinger MM, Kuehl JV, Boore JL, Jansen RK. Genome-wide analyses of Geraniaceae plastid DNA reveal unprecedented patterns of increased nucleotide substitutions. *Proceedings of the National Academy of Sciences.* 2008;105(47):18424–18429.
82. Dugas DV, Hernandez D, Koenen EJ, Schwarz E, Straub S, Hughes CE, et al. Mimosoid legume plastome evolution: IR expansion, tandem repeat expansions, and accelerated rate of evolution in *clpP*. *Sci Rep.* 2015;5:16958.
83. Jiang D. The complete chloroplast genome sequence of *Artemisia stechmaniana* (Asteraceae): genome structure and phylogenetic analysis. *Biologia.* 2024;79(3):715–28.
84. Zhu A, Guo W, Gupta S, Fan W, Mower JP. Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. *New Phytol.* 2016;209(4):1747–56.
85. Xin Y, Yu W, Eiadthong W, Cao Z, Li Q, Yang Z, et al. Comparative analyses of 18 complete chloroplast genomes from Eleven *Mangifera* Species (Anacardiaceae): sequence characteristics and Phylogenomics. *Horticulturae.* 2023;9(1):86.
86. Yengkhom S, Uddin A, Chakraborty S. Deciphering codon usage patterns and evolutionary forces in chloroplast genes of *Camellia sinensis* var. *assamica* and *Camellia sinensis* var. *sinensis* in comparison to *Camellia pubicosta*. *J Integr Agric.* 2019;18(12):2771–85.
87. Xiong AS, Peng RH, Zhuang J, Gao F, Zhu B, Fu XY, et al. Gene duplication, transfer, and evolution in the chloroplast genome. *Biotechnol Adv.* 2009;27(4):340–7.
88. Liu H, Hu H, Zhang S, Jin J, Liang X, Huang B, et al. The complete chloroplast genome of the rare species *Epimedium tianmenshanensis* and comparative analysis with related species. *Physiol Mol Biol Plants.* 2020;26(10):2075–83.
89. Pezosa I, Villacreses J, Rubilar M, Pizarro C, Galleguillos MJ, Ejsmentewicz T et al. Generation of Chloroplast molecular markers to Differentiate *Sophora toromiro* and its hybrids as a First Approach to its reintroduction in Rapa Nui (Easter Island). *Plants.* 2021;10.
90. Zeinalabedini M, Khayam-Nekoui M, Grigorian V, Gradziel TM, Martínez-Gómez P. The origin and dissemination of the cultivated almond as determined by nuclear and chloroplast SSR marker analysis. *Sci Hort.* 2010;125(4):593–601.
91. Geng L, Jiang T, Chen X, Li Q, Ma J, Hou W, et al. Plastome structure, phylogeny and evolution of plastid genes in *Reevesia* (Helicteroideae, Malvaceae). *J Plant Res.* 2024;137(4):589–604.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.