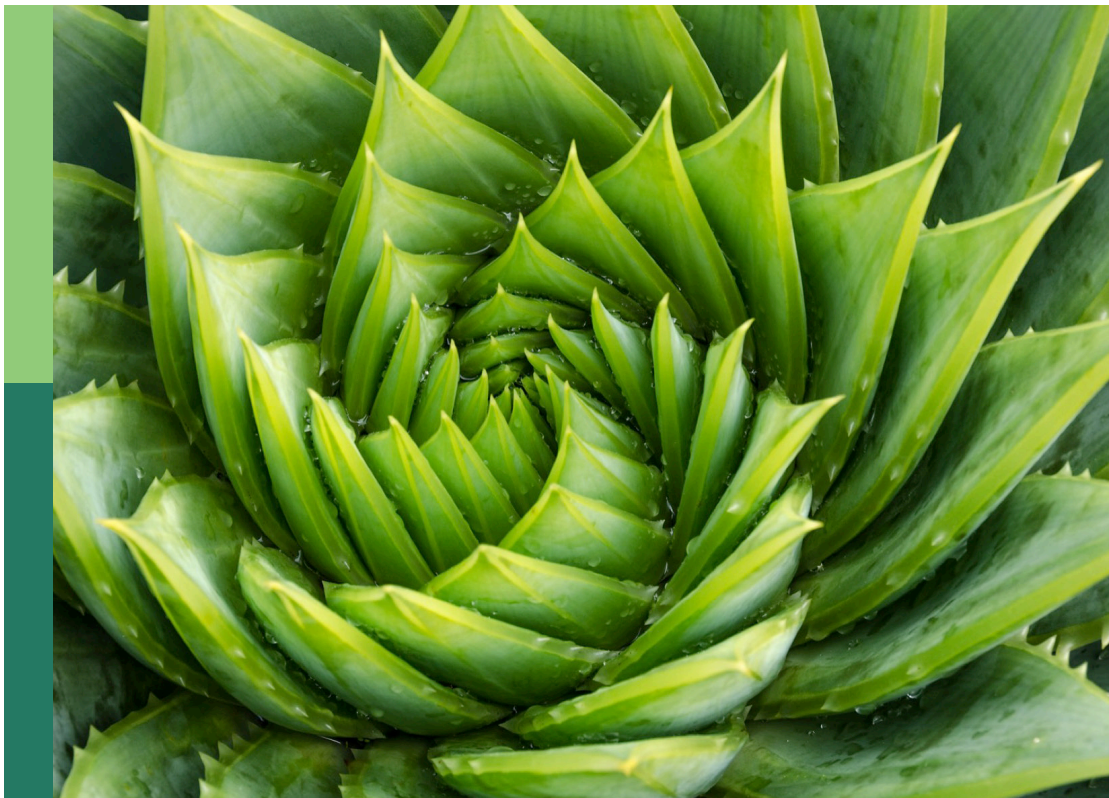
frontiers | Research Topics

Structural variation of the chloroplast genome and related bioinformatics tools

Edited by
Tapan Kumar Mohanta, Baozhong Duan, Weijun Kong, Gang Zhang and Linchun Shi

Published in
Frontiers in Plant Science



ISSN 1664-8714
ISBN 978-2-8325-4811-0
DOI 10.3389/978-2-8325-4811-0

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Structural variation of the chloroplast genome and related bioinformatics tools

Topic editors

Tapan Kumar Mohanta — University of Nizwa, Oman
Baozhong Duan — Dali University, China
Weijun Kong — Capital Medical University, China
Gang Zhang — Shaanxi University of Chinese Medicine, China
Linchun Shi — Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences and Peking Union Medical College, China

Citation

Mohanta, T. K., Duan, B., Kong, W., Zhang, G., Shi, L., eds. (2024). *Structural variation of the chloroplast genome and related bioinformatics tools*.
Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4811-0

Table of contents

frontiers | Frontiers in Plant Science

TYPE Editorial
PUBLISHED 08 April 2024
DOI 10.3389/fpls.2024.1341528

OPEN ACCESS

EDITED AND REVIEWED BY
Wenqin Wang,
Shanghai Normal University, China

*CORRESPONDENCE
Baozhong Duan
bzduan@126.com

RECEIVED 20 November 2023
ACCEPTED 21 March 2024
PUBLISHED 08 April 2024

CITATION
Shi L, Zhang G, Mohanta TK, Kong W and Duan B (2024) Editorial: Structural variation of the chloroplast genome and related bioinformatics tools.
Front. Plant Sci. 15:1341528.
doi: 10.3389/fpls.2024.1341528

Editorial: Structural variation of the chloroplast genome and related bioinformatics tools

Linchun Shi[1], Gang Zhang[2], Tapan Kumar Mohanta[3], Weijun Kong[4] and Baozhong Duan[5]*

[1]State Key Laboratory for Quality Ensurance and Sustainable Use of Dao-di Herbs, Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China, [2]College of Pharmacy, Shaanxi University of Chinese Medicine, Xianyang, Shaanxi, China, [3]Natural and Medical Sciences Research Center, University of Nizwa, Nizwa, Oman, [4]School of Traditional Chinese Medicine, Capital Medical University, Beijing, China, [5]College of Pharmaceutical Science, Dali University, Dali, China

KEYWORDS

chloroplast genome, pseudogenes, gene losss, population study, bioinformatics tools

Editorial on the Research Topic
Structural variation of the chloroplast genome and related bioinformatics tools

1 Introduction

Chloroplast genomes (plastomes) serve as crucial data sources for plant phylogenetic reconstruction and molecular identification. However, comprehensive studies on the assembly, annotation, and in-depth analysis of plastomes exhibiting distinct features such as pseudogenes, gene losses, duplications, rearrangements, widespread intra-individual polymorphisms, and large-scale horizontal gene transfer are still lacking. For instance, the plastomes of various saprophytic and parasitic plants have been observed to undergo a significant loss of photosynthesis-related genes. High-quality assembly, annotation, and in-depth analysis of these plastomes remain major challenges. In addition, more powerful bioinformatics tools are needed to facilitate large-scale plastome studies (Shi et al., 2019). This Research Topic on "Structural Variation of the Chloroplast Genome and Related Bioinformatics Tools" contains ten research articles, emphasizing reliable plastome studies with complex structures or plastomes for population studies, in addition to bioinformatics tools for in-depth analysis of these plastome data. These articles provide a crucial scientific basis for analyzing the structural characteristics of plastomes, studying phylogenetic and genetic evolution, developing new molecular markers, and exploiting bioinformatics tools.

2 Results

2.1 Characterization and in-depth analysis of plastomes with special structures

The plastome of non-photosynthetic plants often undergoes gene loss (Mohanta et al., 2020), pseudogenic transformation, and rearrangement. Orchidacea is a typical taxonomic group, and their plastomes underwent gene loss and gene pseudogenization (Kim et al., 2020). Zhou et al. decoded the plastome of *Galeola lindleyana* in the Vanilloideae subfamily, a heterotrophic orchid plant, and compared it with previously published photosynthetic and heterotrophic orchid plastomes. The length of the *G. lindleyana* plastome has been reduced to 100,749 bp, while still retaining its typical quadripartite structure. In half-heterotrophs, the reduced photosynthetic function begins with the loss of non-essential or stress-related genes, such as *ndh* genes, followed by pseudogenization and the loss of major photosynthesis-related genes (such as *pet*, *psa* and *psb* genes) and plastid-encoded polymerases.

2.2 Large-scale comparative plastome studies for populations

In this Research Topic, there are eight research articles that present comparative analysis of plastomes of plants such as *Solanum lycopersicum* L., *Pseudostellaria heterophylla* (Miq.) Pax, *Corydalis platycarpa* (Maxim. ex Palib.) Makino, *Aristolochia* L., *Phyllanthus* L., *Kandelia* Wight & Arn., *Artemisia* L., and *Paraboea* (Clarke) Ridley. These articles provide valuable information for molecular identification and phylogenetic analysis.

Plastomes can provide distinguishing features and more molecular markers to help identify closely related species and even as super barcodes for accurate identification of plants and germplasm resources (Gao et al., 2023). Lan et al. analyzed 15 plastomes from 5 *Artemisia* species, including 12 newly sequenced genomes. Four hotspot regions and 189-192 SSR molecular markers were identified, which can serve as potential DNA barcodes for further studies on *Artemisia* species. Fang et al. *de novo* assembled and characterized the complete plastomes of nine species of the genus *Phyllanthus*. The authors highlighted three highly variable regions (*trnS-GCU-trnG-UCC*, *trnT-UGU-trnL-UAA*, and *petA-psbJ*) that may be useful as potential molecular markers for identifying *P. urinaria* and its adulterants. Wang et al. sequenced and analyzed the plastomes of 29 tomato germplasm lines. Among the screened SNP markers, those localized to segments of the *ndhH* gene and the *ndhK-ndhC-trnV-UAC* gene spacer region could be used for interspecific identification. The developed SNP markers can be used to analyze genetic diversity and population structure at the plastome level and to develop functional markers associated with traits such as male sterility. Bai et al. sampled 11 species of *Aristolochia* collected from distinct habitats in China and sequenced their complete plastomes. Their analysis of simple sequence repeats (SSRs) was able to identify potential molecular polymorphic markers for analyzing the genetic diversity and structure of *Aristolochia* populations in the future. Highly variable regions would provide candidate markers for *Aristolochia* species identification studies. Zhang et al. collected 17 P. *heterophylla* plant samples with remarkable phenotypic characteristics and obtained their plastome sequences. The authors verified that plastomes could elucidate the relationship among closely related cultivated materials and provide useful information for the development of new, highly polymorphic, and informative molecular makers.

In addition, other articles provide important insights into adaptive evolution and phylogeny, and offer significant guidance for future research on plant evolution and conservation. Raman et al. sequenced the plastome of *C. platycarpa* and conducted wide-scale comparative studies using publicly available data from 20 *Corydalis* plastomes. The results revealed extensive genome rearrangement and IR expansion, events that evolved independently in the *Corydalis* species. The divergence time of the *ndh* gene in the *Corydalis* sub-clade species (44.31-15.71 mya) is consistent with the uplift of the Qinghai-Tibet Plateau in the Oligocene and Miocene, and may have triggered the radiation of the *Corydalis* species during this period. In 2003, *Kandelia obovata* was identified as a new mangrove species distinct from *Kandelia candel*. Xu et al. sequenced the 25 whole plastomes of *K. obovata* (18 samples) and *K. candel* (Seven samples) for comparison. A comparative molecular simulation study of the homologous NAD(P)H dehydrogenase chain 4 (NDH-D) and ATP synthase subunit alpha (ATP-A) proteins of *K. candel* and *K. obovata* predicted that the functions of photosynthetic electron transport and ATP generation were significantly different. The results suggest that energy demand is a pivotal factor in their adaptation to different environments geographically separated by the South China Sea. Wang et al. sequenced and compared the complete plastomes of 12 *Paraboea* species from China and Vietnam. The study presents several important findings:

1. It demonstrates the strong conservation of the plastomes among *Paraboea* species.
2. It confirms the monophyletic nature of the genus. The study also reveals the impact of purifying selection on the protein-coding genes in the plastomes.
3. It emphasizes the significance of understanding the genetic mechanisms underlying plant adaptation to specific environmental conditions, particularly karst environments.

2.3 Bioinformatics tools for plastome annotation and analysis

Genome comparison of multiple plastomes involves many software, including some popular tools, such as mVISTA, Mauve, IRscope, etc. However, these tools are separate from each other and lack a unifying interface, making comparison analysis a time-consuming and labor-intensive task. An automated workflow for fast and accurate assembly and annotation of plastome sequences from raw whole (nuclear) genome sequencing data is needed. Chen et al. developed Plastaumatic-an automated pipeline for both the assembly and annotation of plastomes, with the ability to load

whole genome sequence data with minimal manual input and, therefore a faster run time. The pipeline was demonstrated on two sets of plant sequence data: three soybean accessions and 12 potato accessions. It showed substantially faster completion than manual assembly. This automated pipeline, which includes plastid assembly and annotation is an efficient tool. In their article, Zhou et al., used the GetOrganelle (V1.7.7.0) assembly toolkit which provides fast and accurate *de novo* assembly of the organelle genome. They also used the CPGAVAS2 web server, which automatically annotates the plastome.

3 Perspectives

This Research Topic has conducted a series of comparative studies on plastomes, with their characterization and in-depth analysis, and on bioinformatics analysis tools. These contributions are of great value for the analysis of genomic features, phylogenetic and genetic evolution, and the development of new molecular markers. However, more research is needed on the plastomes of parasitic plants, saprophytic plants, or plants living under extreme environmental conditions, such as drought, cold, high altitude, and high soil salinity environments. For instance, during the rapid radiation and evolution of *Rhodiola* plants growing at high altitudes, the evolution of plant plastomes and their role in adaptation to extreme habitats remained largely unknown. Therefore, the study of plant plastomes under harsh environmental conditions emerges as an important research direction for the future, which will help deepen our understanding of the mechanisms behind plant adaptation to extreme environments and will provide crucial insights for future studies on plant evolution and conservation.

In addition, the plant chloroplast genome has reached a critical point. With recent advances in sequencing technology, the speed of generating chloroplast genomes has increased dramatically. The number of chloroplast genome sequences of Viridiplantae plants has increased from about 3,000 in 2019 to more than 25,000 currently. However, there are several fundamental issues that need to be addressed urgently: i, A standardized nomenclature for tRNA genes and protein-coding genes has not yet been uniformly accepted for all chloroplast genomes due to the lack of a plastid gene nomenclature committee like that for humans (https://www.genenames.org/), and this has prevented the large-scale comparative analysis of chloroplast genome data in an automated way. The proposed solution to this issue may start with the naming uniformity among several popular annotation tools, such as Geseq, CPGAVAS2, AGORA, etc., or by public databases, such as NCBI. ii, The number of chloroplast coding genes may be obscured in some taxonomic groups due to over-reliance on chloroplast genome annotation tools. The two popular annotation tools, GeSeq and CPGAVAS2, rely on a curated reference sequence database, and *de novo* annotation tools are still lacking. iii, The methods of comparative analysis of chloroplast genomes tend to be similar, and new essential tools need to be developed to facilitate the deep mining of chloroplast genome data, such as species differentiation, species expansion, and species adaptation to the extreme environment.

Author contributions

LS: Writing – review & editing, Writing – original draft. GZ: Writing – review & editing. TM: Writing – review & editing, Writing – original draft. WK: Writing – review & editing. BD: Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was supported by the Key project at the central government level: The ability establishment of sustainable use for valuable Chinese medicine resources (No. 2060302).

Acknowledgments

We thank all the authors and the reviewers who contributed to this Research Topic.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

References

Gao, M., Huo, X., Lu, L., Liu, M., and Zhang, G. (2023). Analysis of codon usage patterns in Bupleurum falcatum chloroplast genome. *Chin. Herb Med.* 15 (2), 284–290. doi: 10.1016/j.chmed.2022.08.007

Kim, Y. K., Jo, S., Cheon, S. H., Joo, M. J., Hong, J. R., Kwak, M., et al (2020). Plastome evolution and phylogeny of orchidaceae, with 24 new sequences. *Front. Plant Sci.* 11, 22. doi: 10.3389/fpls.2020.00022

Mohanta, T. K., Mishra, A. K., Khan, A., Hashem, A., Abd_Allah, E. F., and Al-Harrasi, A. (2020). Gene loss and evolution of the plastome. *Genes.* 11 (10), 1133. doi: 10.3390/genes11101133

Shi, L. C., Chen, H. M., Jiang, M., Wang, L. Q., Wu, X., Huang, L. F., et al. (2019). CPGAVAS2, an integrated plastome sequence annotator and analyzer. *Nucleic Acids Res.* 47 (W1), W65–W73. doi: 10.1093/nar/gkz345

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 05 October 2022
DOI 10.3389/fpls.2022.1019831

OPEN ACCESS

EDITED BY
Weijun Kong,
Capital Medical University, China

REVIEWED BY
Hui Yao,
Chinese Academy of Medical Sciences and Peking Union Medical College, China
Ramasamy Yasodha,
ICFRE, India
Weichao Ren,
Heilongjiang University of Chinese Medicine, China

*CORRESPONDENCE
Bo Zhao
122017017@glmc.edu.cn
Yaolei Mi
xiaomi20063@sina.com

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

RECEIVED 15 August 2022
ACCEPTED 16 September 2022
PUBLISHED 05 October 2022

CITATION
Wang Y, Wen F, Hong X, Li Z, Mi Y and Zhao B (2022) Comparative chloroplast genome analyses of *Paraboea* (Gesneriaceae): Insights into adaptive evolution and phylogenetic analysis.
Front. Plant Sci. 13:1019831.
doi: 10.3389/fpls.2022.1019831

Comparative chloroplast genome analyses of *Paraboea* (Gesneriaceae): Insights into adaptive evolution and phylogenetic analysis

Yifei Wang[1,2†], Fang Wen[3†], Xin Hong[4,5], Zhenglong Li[4,5], Yaolei Mi[6*] and Bo Zhao[1,2,3*]

[1]Department of Pharmacognosy, Guilin Medical University, Guilin, China, [2]Department of Pharmacy, Guilin Medical University, Guilin, China, [3]Guangxi Key Laboratory of Plant Conservation and Restoration Ecology in Karst Terrain, Guangxi Institute of Botany, Guangxi Zhuang Autonomous Region and Chinese Academy of Sciences, Guilin, China, [4]Anhui Provincial Engineering Laboratory of Wetland Ecosystem Protection and Restoration, School of Resources and Environmental Engineering, Anhui University, Hefei, China, [5]Yunnan Key Laboratory for Integrative Conservation of Plant Species with Extremely Small Populations, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, China, [6]Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing, China

Paraboea (Gesneriaceae) distributed in the karst areas of South and Southwest China and Southeast Asia, is an ideal genus to study the phylogeny and adaptive evolution of karst plants. In this study, the complete chloroplast genomes of twelve *Paraboea* species were sequenced and analyzed. Twelve chloroplast genomes ranged in size from 153166 to 154245 bp. Each chloroplast genome had a typical quartile structure, and relatively conserved type and number of gene components, including 131 genes which are composed of 87 protein coding genes, 36 transfer RNAs and 8 ribosomal RNAs. A total of 600 simple sequence repeats and 389 non-overlapped sequence repeats were obtained from the twelve *Paraboea* chloroplast genomes. We found ten divergent regions (*trn*H-GUG-*psb*A, *trn*M-CAU, *trn*C-GCA, *atp*F-*atp*H, *ycf1*, *trn*K-UUU-*rps*16, *rps*15, *pet*L, *trn*S-GCU-*trn*R-UCU and *psa*J-*rpl*33) among the 12 *Paraboea* species to be potential molecular markers. In the phylogenetic tree of 31 Gesneriaceae plants including twelve *Paraboea* species, all *Paraboea* species clustered in a clade and confirmed the monophyly of *Paraboea*. Nine genes with positive selection sites were detected, most of which were related to photosynthesis and protein synthesis, and might played crucial roles in the adaptability of *Paraboea* to diverse karst environments. These findings are valuable for further study of the phylogeny and karst adaptability of Gesneriaceae plants.

KEYWORDS

***Paraboea*, phylogenetic, positive selection, chloroplast, genome**

Introduction

Limestone areas have diverse and unique regional microhabitat and special island habitats (Xin et al., 2021). These special habitats (such as stone mountains, karst caves, Tiankeng, etc.) provide very favorable conditions for species isolation and differentiation. After a long period of evolution and natural selection, limestone areas have bred a high degree of species diversity and significant endemism. With the development of the economy, due to frequent human activities (such as tourism development, etc.), the fragile ecological balance of these diversified microhabitats is easily destroyed, and the survival of karst plants is threatened. Studying the adaptive evolution of karst plants has extremely important practical significance for the protection of karst plants (Tao et al., 2016).

Paraboea, an important genus of Gesneriaceae, contains about 144 species (Xu et al., 2012; Wen et al., 2013; Puglisi and Phutthai, 2017). The genus is mainly distributed in karst areas of south China, southwest China and southeast Asia, China has about 29 species, of which 19 are endemic (Guo et al., 2016). Among these species, some are widely distributed and attached to the rock gaps of stone mountains in direct sunlight (such as *P. rufescens* and *P. swinhoei*), some are distributed in dark and wet caves (such as *P. filipes*), and some grow on limestone rocks or inter stone soil in the dark place under the dense forest from the hillside to the top of the mountain (such as *P. barbatipes*) (Zhou et al., 2003; Gao et al., 2006). Considering the distribution of *Paraboea* species in diverse microhabitats, it becomes an excellent group to study the adaptive evolution of karst plants. However, the phylogenetic study of *Paraboea* is mainly based on a small number of chloroplast or nuclear gene markers, and the phylogenetic relationship has not been completely solved, which greatly limits the discussion of adaptive evolution of genes and traits. Based on nuclear ITS (internal transcribed spacer) sequences and chloroplast genome sequences (*trn*H-*psb*A spacer), the phylogenetic relationship of *Paraboea* was reconstructed, and the taxonomic boundaries among some related species were clarified (Li and Wang, 2007; Puglisi et al., 2011; Xin et al., 2019; Guo et al., 2020). However, the existing chloroplast genome sequences cannot completely solve the phylogenetic relationship of *Paraboea*. It is necessary to add faster and more suitable molecular sequences to reconstruct its phylogenetic relationship.

The chloroplast genome is one of the important molecular tools to study plant adaptive evolution. The challenging environment may impose selective pressure on genes related to photosynthesis, leaving the footprints of natural selection on genes. The main protein coding genes of the chloroplast genome include those controlling genetic and photosynthetic systems as well as genes encoding other functions. Photosynthetic system genes are genes related to photosynthesis, which are responsible for encoding members of ATP synthase, Rubisco large subunit, NADPH dehydrogenase and photosystem I and II (Zhang et al., 2018). The adaptive evolution analysis showed that chloroplast genes related to photosynthesis generally had positive selection sites in plants living in various extreme environments, and these gene regions might play a crucial role in plant adaptation to different environments (Chen et al., 2021).

The chloroplast genome sequences not only provide full-length protein coding sequences for the adaptive evolution of genes related to photosynthesis under the selection pressure of different environments, but also screen suitable hypervariable regions to solve the phylogenetic relationship of plants (Yang et al., 2020). The size of chloroplast genome in terrestrial plants is 120-160 kb, encoding 110-130 unique genes. Because of the slow evolutionary rate of change, maternal inheritance, less recombination and satisfactory collinearity between the sequences of various plant groups, the chloroplast genome sequences were suitable for molecular markers (Zhai et al., 2021). With the development of Next-generation sequencing technology, a large number of chloroplast genome data can be easily obtained. Based on chloroplast comparative genomics analyses, the high variation regions were located to develop specific molecular markers of groups or species for applying to the research of phylogenetic analysis and species identification (Chen et al., 2022; Song et al., 2022).

So far, there was no scientific research related to the complete chloroplast genome of *Paraboea*. In this study, we sequenced, assembled and analyzed the chloroplast genomes of twelve *Paraboea* species, and constructed the phylogenetic relationship of 31 species belonging to 12 genera of Gesneriaceae based on protein coding sequences. We also calculated selective pressures to investigate whether the coding protein genes in *Paraboea* species were under purifying selection or positive selection. Comprehensive insights into the character and evolution of the chloroplast genomes, provided a theoretical basis for the protection and rational utilization of germplasm resources of *Paraboea* plants in karst areas.

Materials and methods

Plant materials and DNA extraction

The 12 species of *Paraboea* in China and Vietnam used in the study were identified, collected and finally cultivated in the Guangxi Institute of Botany (Table 1). Fresh green leaves were sampled, washed, dried and stored at -80°C till DNA extraction (Feng et al., 2020). The total genomic DNA was extracted according to the modified CTAB method (Doyle and Doyle, 1987).

TABLE 1 Sources of material from twelve *Paraboea* species.

Taxon	Voucher	GenBank accession number	Location	Habitat
P. clavisepala	ZBPC202100061	MZ465381	Jingxi Guangxi, China	Limestone; ca. 800 m
P. dictyoneura	ZBPD202100062	MZ465383	Yingde Guangdong, China	Rocks in forests; 100-800 m
P. dolomitica	ZBPD202100063	MZ465376	Shibing Guizhou, China	rock faces of dolomite karst area, ca. 650-855 m
P. filipes	ZBPF202100064	MZ465379	Lianzhou Guangdong, China	Limestone cliffs; ca.100-300 m
P. glutinosa	ZBPG202100065	MZ465382	Caobang, Vietnam	Rocks of slopes; ca. 400-1400 m
P. guilinensis	ZBPG202100066	MZ465377	Guilin Guangxi, China	Limestone cliffs
P. martinii	ZBPM202100067	MZ465385	Napo Guangxi, China	limestone under the hillside forest; ca. 1220-1260 m
P. peltifolia	ZBPP202100068	MZ465386	Mashan Guangxi, China	Limestone; ca. 300-400 m
P. rufescens	ZBPR202100069	MZ465384	Napo Guangxi, China	On rocks of limestone hills and valley forests; ca. 200-1500 m
P. sinensis	ZBPS202100070	MZ465380	Longzhou Guangxi, China	Crevices of rocks or on cliffs in forests; ca. 600-2500 m
P. swinhoei	ZBPS202100071	MZ465378	Rongshui Guangxi, China	Shady and damp rocks under forests; ca. 300-1000 m
P. wenshanensis	ZBPW202100072	MZ465375	Wenshan Yunnan, China	moist shady cliffs of limestone hills, ca. 1500 m

Genome sequencing and assembling

Qualified DNA fragments were obtained by mechanical fracture method, and were sequenced after purification, terminal repair and other processing. The 350 bp fragment was screened by agarose gel electrophoresis and amplified by PCR to construct a sequence library. Paired-end (PE) reads were obtained using the Illumina HiSeq 2000 sequencer (Illumina Biotechnology Company, San Diego, CA, USA) (Gu et al., 2018). *De novo* genome assembly from the clean data was accomplished utilizing NOVOPlasty v2.7.2 (Dierckxsens et al., 2017), with a k-mer length of 39 bp and the chloroplast genome of *Primulina huaijiensis* (NC_036413) as the reference sequence. The correctness of the assembly was confirmed by manually editing and mapping all the raw reads to the assembled genome sequence using Bowtie2 (v2.0.1) (Langmead et al., 2009) under the default settings. Finally, the complete chloroplast genome sequences of twelve *Paraboea* species were obtained.

Genome annotation and sequence characterization

Functional annotation of the chloroplast genome includes coding gene prediction and non-coding RNA (rRNA and tRNA) annotation. Using CPGAVAS2 (Shi et al., 2019), the twelve complete chloroplast genomes were annotated with a reference genome (*Primulina huaijiensis*, GenBank: NC036413). Meanwhile, tRNA scan-SE version 1.21 (Schattner et al., 2005) was used to identify and confirm tRNA genes. The twelve circular chloroplast genome maps were constructed using the OrganellarGenomeDRAW (OGDRAW) v.1.3.1 tool followed by manual modification (Greiner et al., 2019). And the whole twelve sequences were submitted to GenBank (Table 1).

Repeat sequences and SSR analysis

The Perl script MISA (http://pgrc.ipk-gatersleben.de/misa/) (Beier et al., 2017) was used with the filter thresholds set to detect SSRs. The specific parameters were set at repeat units ≥ 8 for mononucleotides, repeat units ≥ 4 for dinucleotides and trinucleotides, and repeat units ≥ 3 for tetranucleotides, pentanucleotides and hexanucleotides. To identify complex repeative sequences such as forward, reverse, complement and palindromic, REPuter online software (Kurtz et al., 2001) was used with a minimum repeat size of 30 bp and 90% sequence identity (Hamming distance of 3).

Boundary regions and genome comparative analysis

In order to better display the expansion/contraction events of the IR region, the connecting regions of IR-LSC and IR-SSC in the chloroplast genomes of twelve *Paraboea* species were compared by using IRscope (https://irscope.shinyapps.io/irapp/) online software (Amiryousefi et al., 2018). To identify interspecific variations, the mVISTA online software was used to compare the chloroplast genomics of twelve *Paraboea* plants (Frazer et al., 2004). The comparative analysis was carried out by using the shuffle-LAGAN mode in mVISTA with the annotation of *P. sinensis* as reference, and the sequence alignment was visualized in an mVISTA plot. We used MEGA v6.0 (Tamura et al., 2013) to calculate the percentage of variable sites in the protein-coding genes. We also used DnaSP v6.0 (Rozas et al., 2017) to calculate the nucleotide polymorphism (Pi) among the twelve *Paraboea* species. When calculating the Pi value, set the windows length to 100 sites and the step size to 25 sites.

Phylogenetic analysis

The complete chloroplast protein-coding genes of 31 Gesneriaceae species (12 *Paraboea* species in this study and 20 other species from NCBI) were aligned using MUSCLE v3.8.31 (Edgar, 2004), and then aligned in MAFFT (version 7.222) using the default parameters (Kazutaka and Standley, 2013). The final two sequence alignment results are consistent. The aligned sequences were used to construct the phylogenetic trees using the maximum likelihood (ML) method implemented in RAxML 7.0.4 (Stamatakis, 2006) with 1000 replicates under the GTR + CAT model.

Adaptive evolution analysis

In order to detect the positive selection of chloroplast genes in *Paraboea*, the non-synonymous (DN) and synonymous (DS) substitution rates of protein-coding genes and the DN/DS (ω) values of protein-coding genes were calculated. All of the CDS sequences were extracted from chloroplast genome, and then the single-copy CDS sequences common to all species were selected and aligned with the codon model. We used EasyCodeML v1.21 (Gao et al., 2019b) to identified positive selection sites. A total of 76 CDSs presented in all the analysed species, and were used for identification of positive selection using the site model (seqtype = 1, model = 0, NSsites = 0, 1, 2, 3, 7, 8). In addition, Bayes Empirical Bayes (BEB) method (Huelsenbeck and Ronquist, 2001) was used to calculate the posterior probabilities for amino acid sites that were potentially under positive selection. The results showed that the amino acid sites with a posteriori probability of more than 0.95 were positive selected. Moreover, the logarithmic likelihood value of site models was calculated by likelihood ratio test (LRT) and its statistical significance. Finally, we used the PSIPRED server (Buchan et al., 2013) to visualize the amino acid sequences of positively selected gene secondary structure, and used the SWISS-MODEL online software (Waterhouse et al., 2018) to predict the protein structure of these genes.

Results

General features of chloroplast genomes

In this study, the chloroplast genomes of twelve *Paraboea* species were sequenced and characterized. Each chloroplast genome was made up of three distinct regions: a small single copy region (SSC), a large single copy region (LSC) and two inverted repeat regions (IRs) (Figure 1). The complete chloroplast genomes of the 12 *Paraboea* species ranged from 153166 bp (*P. guilinensis*) to 154245 bp (*P. wenshanensis*) in length (Table 2). The length of SSC ranged from 17656 bp (*P. glutinosa*) to 18089 bp (*P. wenshanensis*), while the length of LSC and IR length ranged from 84761 bp (*P. clavisepala*) to 85488 bp (*P. wenshanensis*), and from 25272 bp (*P. dolomitica*) to 25334 bp (*P. wenshanensis*). In all twelve *Paraboea* species, the chloroplast genomes of *P. filipes* and *P. wenshanensis* had the lowest total GC content (37.45%), while the chloroplast genome of *P. martini*had the highest total GC content (37.72%). Gene annotation showed that each chloroplast genome contained 131 genes in conserved order and orientation, which contained 8 ribosomal RNA (rRNA) genes, 36 transfer RNAs (tRNAs), and 87 protein-coding genes (Table 3). Fifteen genes (10 protein coding genes and 5 tRNA genes) with introns were identified. Among them, the *clp*P and *ycf3* genes had two introns, respectively, while the other 13 genes had one intron.

IR expansion and contraction in the twelve *Paraboea* chloroplast genomes

There were 4 borders between LSC, IRb, IRa and SSC in the cpGenome: LSC/IRb border (JLB line), IRb/SSC border (JSB line), SSC/IRa border (JSA line), IRa/LSC border (JLA line). The borders of the twelve *Paraboea* chloroplast genomes were compared (Figure 2). The LSC/IRb border and IRa/LSC border were relatively conservative. The *rpl2* gene located at the LSC/IR border, and the distances between *rps2* and the JLB line ranged from 41 bp to 95 bp. The *trn*H-GUG noncoding gene located on the right side of the JLA line with a distance of 0 to 9 bp.

At the IRb/SSC border, the *ndh*F encoding gene located at the IRB-SSC boundary. In the chloroplast genome of *P. sinensis*, *P. swinhoei*, *P. peltifolia*, *P. filipes* and *P. guilinensis*, the *ndh*F gene had the length of 72 bp (*P. guilinensis*) to 138 bp (*P. sinensis*) in the IRB region. In the other seven *Paraboea* chloroplast genomes, the *ndh*F gene spanned the IRb/SSC border and had the length of 113 bp to 124 bp.

At the SSC/IRa border, the *ycf1* gene spanned the SSC-IRB boundary. Due to the special position of the *ycf1* gene, there were seven *Paraboea* chloroplast genomes in the IRA region with *ycf1* pseudogenes, the corresponding length ranged from 842 bp to 863 bp in the IRA region. And in the other *Paraboea* species chloroplast genomes, the *ycf1* gene was located on the SSC-IRA boundary, which made the corresponding pseudogene take place in the IRB region with the length of 799 bp to 833 bp.

Repeat sequence analysis

The twelve *Paraboea* chloroplast genomes contained 600 SSRs (Figure 3A and Supplementary Table S2). In the chloroplast genome of *P. rufescens*, 41 SSRs were detected, which was the least of the 12 chloroplast genomes. And in the chloroplast genome of *P. sinensis*, a total of 57 SSRs were identified, which was the most of the 12 chloroplast genomes. For each *Paraboea* specie, mononucleotide repeats were the

10.3389/fpls.2022.1019831

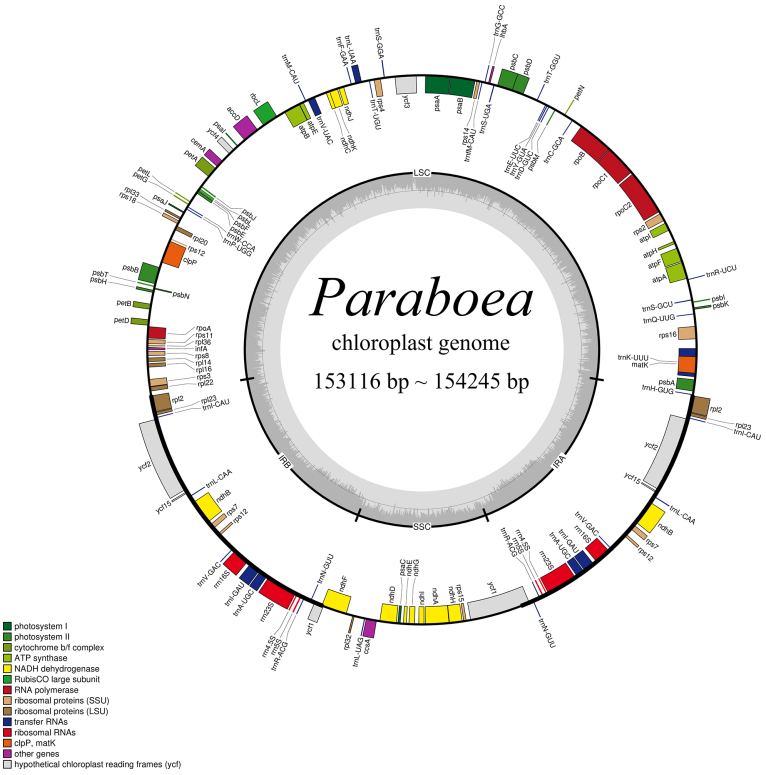


FIGURE 1
Structural map of the *Paraboea* chloroplast genome. Genes shown outside the outer circle are transcribed clockwise and those inside are transcribed counterclockwise. Genes belonging to different functional groups are color-coded. *P. clavisepala* is used as the template for Figure 1.

most common, with numbers ranging from 19 to 34; followed by tetranucleotides ranging from 10 to 16; dinucleotides ranging from 6 to 14; trinucleotides ranging from 1 to 4; pentanucleotides ranging from 0 to 2 and hexanucleotide ranging from 0 to 2 (Supplementary Figure S1).

Non-overlapped sequence repeats including forward repeats, reverse repeats, palindromic repeats and complement repeats were detected in twelve chloroplast genomes. A total of 389 non-overlapped sequence repeats were detected in twelve chloroplast genomes of *Paraboea* plants (Figure 3B and Supplementary

TABLE 2 Summary of the chloroplast genomes of twelve *Paraboea* species.

	Genome Length (bp)	LSC Length (bp)	SSC Length (bp)	IR Length (bp)	GC (%)	Total Genes	CDS	tRNA	rRNA
P. clavisepala	153398	84761	18045	25296	37.58%	131	87	36	8
P. dictyoneura	153406	84829	17999	25289	37.52%	131	87	36	8
P. dolomitica	153510	84885	18081	25272	37.45%	131	87	36	8
P. filipes	153486	84851	18001	25317	37.45%	131	87	36	8
P. glutinosa	153505	85303	17656	25273	37.66%	131	87	36	8
P. guilinensis	153166	84819	17759	25294	37.57%	131	87	36	8
P. martinii	153580	84978	18032	25285	37.72%	131	87	36	8
P. peltifolia	153459	84784	18043	25316	37.55%	131	87	36	8
P. rufescens	153352	85098	17694	25280	37.69%	131	87	36	8
P. sinensis	153453	84869	18028	25278	37.61%	131	87	36	8
P. swinhoei	153564	85160	17800	25302	37.65%	131	87	36	8
P. wenshanensis	154245	85488	18089	25334	37.70%	131	87	36	8

LSC, large single-copy; SSC, small single-copy; IR, inverted repeat.

10.3389/fpls.2022.1019831

TABLE 3 Genes in the chloroplast genome of twelve *Paraboea* species.

Category	Gene group	Gene name
Protein synthesis and DNA-replication	Ribosomal RNA genes	*rrn4.5, rrn5, rrn16, rrn23*
	Transfer RNA genes	*trnA-UGC*, trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnG-GCC, trnH-GUG, trnI-CAU, trnI-GAU*, trnK-UUU*, trnL-CAA, trnL-UAA*, trnL-UAG, trnM-CAU, trnfM-CAU, trnN-GUU, trnP-UGG, trnQ-UUG, trnR-UCU, trnR-ACG, trnS-GCU, trnS-GGA, trnS-UGA, trnT-GGU, trnT-UGU, trnV-GAC, trnV-UAC*, trnW-CCA, trnY-GUA*
	Ribosomal protein genes (larger subunit)	*rpl2*, rpl14, rpl16*, rpl20, rpl22, rpl23, rpl32, rpl33, rpl36*
	Ribosomal protein genes (smaller subunit)	*rps2, rps3, rps4, rps7, rps8, rps11, rps12, rps14, rps15, rps16, rps18*
	RNA polymerase	*rpoA, rpoB, rpoC1*, rpoC2*
Photosynthesis	Photosystem I	*psaA, psaB, psaC, psaI, psaJ*
	Photosystem II	*psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT*
	Cytochrome b/f complex	*petA, petB, petD, petG, petL, petN*
	ATP synthase	*atpA, atpB, atpE, atpF*, atpH, atpI*
	Rubisco large subunit	*rbcL*
	NADH dehydrogenase	*ndhA*, ndhB*, ndhC, ndhD, ndhE, ndhF*, ndhG, ndhH, ndhI, ndhJ, ndhK*
Miscellaneous group	ATP-dependent protease	*clpP***
	Maturase	*matK*
	Acetyl-CoA carboxylase	*accD*
	Cytochrome c biogenesis	*ccsA*
	Inner membrane protein	*cemA*
Pseudogene unknown function	Hypothetical chloroplast reading frames (ycf)	*ycf1*, ycf2, ycf3**, ycf4, ycf15*
Other gene	LhbA	*lhbA*

"*" indicates the presence of one intron.
"**" indicates the presence of two introns.

Table S3). The number of non-overlapped sequence repeats varied from 28 in *P. dictyoneura* to 37 in *P. sinensis*. Among these non-overlapped repeats, palindromic repeats were the most common with 207, followed by forward repeats with 160; reverse repeats with 15 and complement repeats with 7 (Supplementary Figure S2). The repeat sequence analysis would provide help for the study of genetic variation in *Paraboea*.

Comparative chloroplast genome analysis

Taking *P. sinensis* as a reference, multiple alignments of twelve *Paraboea* chloroplast genomes were conducted, and the results suggested that the non-coding sequences showed more divergence than the coding regions (Figure 4). According to the comparative analysis, the main divergent sequences for the noncoding regions were *atp*H-*atp*I, *atp*F-*atp*H, *rps16*-*trn*Q-UUG, *trn*K-UUU-*rps16*, *trn*H-GUG-*psb*A, *trn*S-GCU-*trn*R-UCU and *psa*A-*ycf3*, and the strongly divergent sequences for the coding regions were *mat*K, *pet*L, *ycf1*, *ycf2* and *ndh*F, which might be good candidates for *Paraboea* species identification. To quantify the levels of DNA polymorphism, we calculated the Pi values of above twelve regions, the Pi values of these regions were calculated ranged from 0.01569 (*psa*A-*ycf3*) to 0.08362 (*trn*H-GUG-*psb*A) (Figure 5A). The highest average Pi value of the coding regions was calculated in the SSC region, followed by the coding regions of the LSC and IR region (Figures 5B–D). The ten coding genes with the highest polymorphism in descending order include: *ycf1*, *rps15*, *pet*L, *mat*K, *rpl22*, *ndh*F, *rps3*, *rps8*, *psa*I and *ccs*A. The Pi values of tRNA and rRNA genes were also

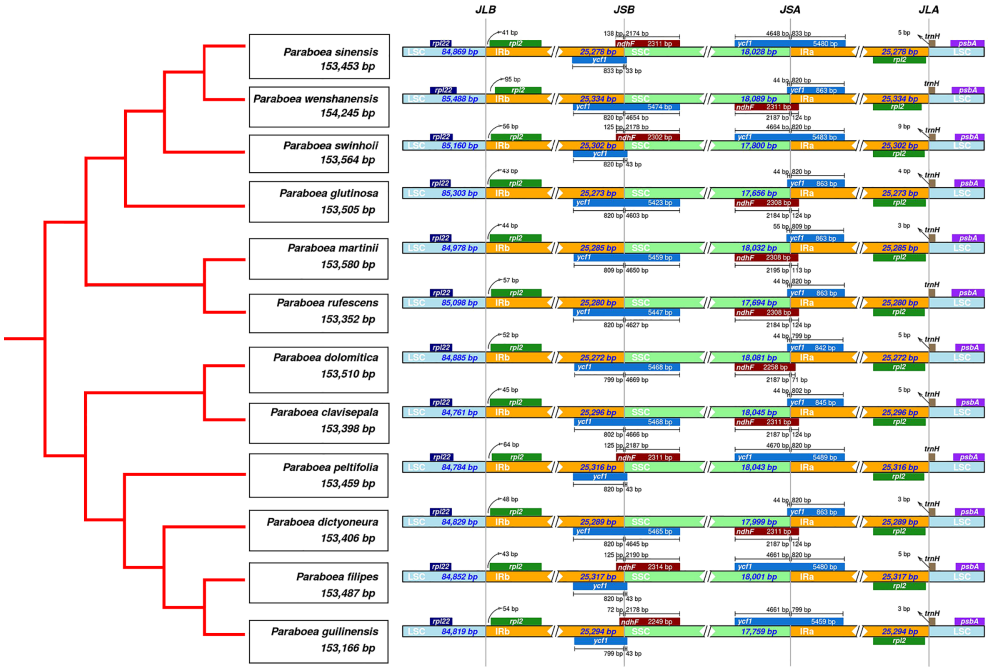


FIGURE 2
Comparison of the borders of the LSC, SSC, and IR regions among twelve chloroplast genomes.

calculated, and the results showed that *trn*C-GCA and *trn*M-CAU had high Pi values, 0.0676 and 0.068, respectively. We finally screened out ten divergent regions with the highest value, which were *trn*H-GUG-*psb*A, *trn*M-CAU, *trn*C-GCA, *atp*F-*atp*H, *ycf1*, *trn*K-UUU-*rps16*, *rps15*, *pet*L, *trn*S-GCU-*trn*R-UCU and *psa*J-*rpl33*. These divergent regions may be the best candidate marker for DNA barcoding.

Phylogenetic relationship

In order to study the phylogenetic position of *Paraboea*, ML tree were constructed using 76 protein coding genes of the chloroplast genomes for 31 Gesneriaceae species, including 12 *Paraboea* species (Figure 6). Among the 31 Gesneriaceae species, except for 12 *Paraboea* species, the chloroplast genomes of the remaining species were obtained from NCBI (Supplementary Table S1). In the phylogenetic tree, all nodes were supported with bootstrap values greater than 60%, and each genus clustered together into a clade (100% bootstrap values). The 12 *Paraboea* species clustered into a clade, and then clustered with *Dorcoceras hygrometrica* (100% bootstrap values). *Paraboea* clade were divided into two major small clades with 100% bootstrap support value. In one major small clade, *P. clavisepala* and *P. dolomitica* form a clade, and then sequentially formed clades with *P. peltifolia*, *P. dictyoneura*, *P. guilinensis* and *P. filipes*. In another major small clade, the clade formed by *P. sinensis* and *P. wenshanensis*, clustered with *P. rufescens* and *P. glutinosa*, and then shared a sister relationship with *P. martinii* and *P. swinhoei*.

Adaptive evolution analysis

The 76 chloroplast protein coding genes of twelve *Paraboea* species were tested, and positive selection was found in nine genes (*lhb*A, *mat*K, *ndh*F, *psb*K, *rbc*L, *rpl22*, *rps12*, *rps18* and *ycf1*) with a high posterior probability (>95%) using the BEB test (Figure 7 and Supplementary Table S4). One amino acid site (the 39th codon) was identified to be under positive selection in *lhb*A gene (Figure 7A). The spatial analysis of LhbA protein under positive selection indicated that the site was located in the α-helix (Figure 8A). Four amino acid sites (the 81th, 116th, 284th and 353th codons) under positive selection were detected in Maturase coded by *mat*K gene (Figure 7B). The spatial analysis indicated that two sites were located in α-helix, and the other sites were located at β-sheet and random coil, respectively (Figure 8B). In addition, three of nine genes were coding genes for photosynthesis: the *ndh*F gene for NADH dehydrogenase subunit F (NDHF), the *psb*K gene for Photosystem II subunits K (PsbK), and the *rbc*L gene for rubisco large subunit (RBCL). Three amino acid sites (463th, 651th and 729th) under positive selection in NDHF were located the random coil, α-helix and α-helix respectively (Figure 7C and Supplementary Figure S3). Based on the protein structure prediction, one amino acid site

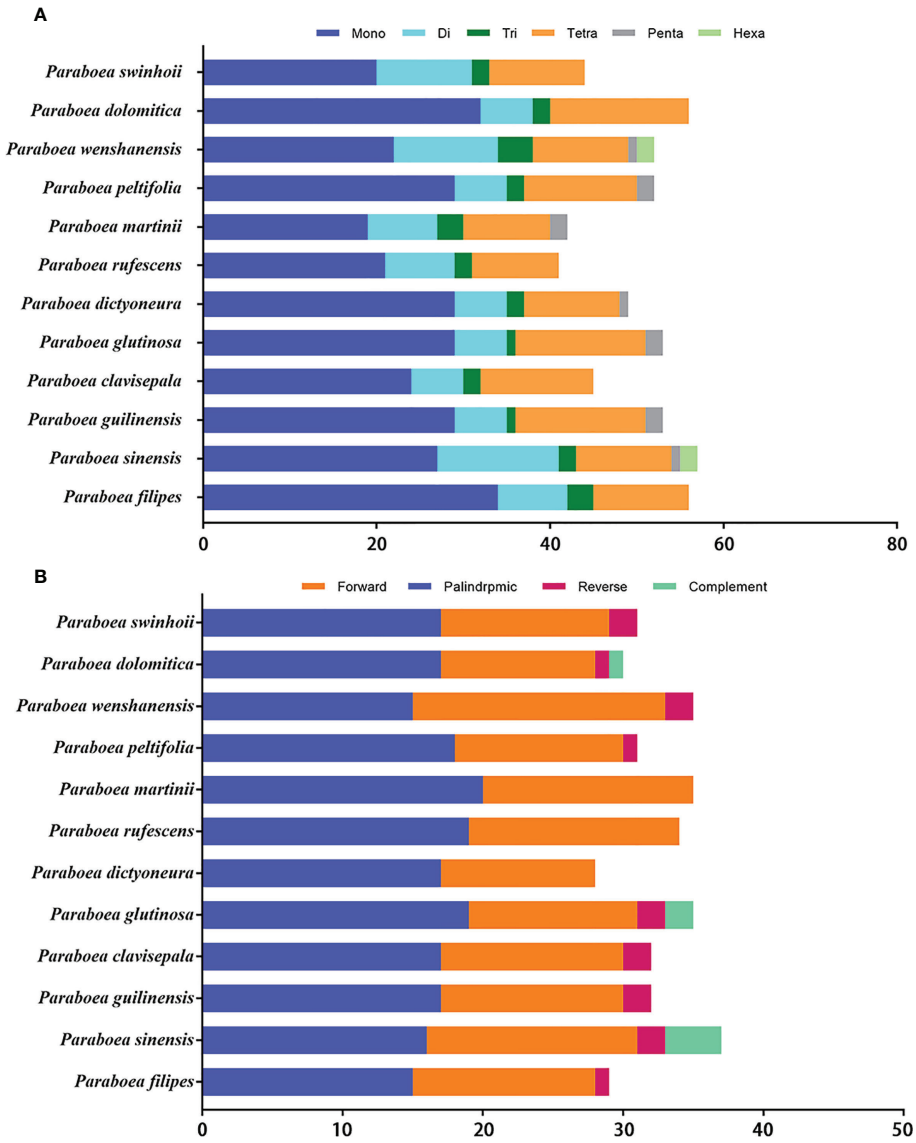


FIGURE 3
The distribution maps of sequence repeats. **(A)** Number of different types of SSRs present in twelve *Paraboea* chloroplast genomes. **(B)** The comparison of the four complex repeat types among twelve *Paraboea* choloplast genomes.

(34th) under positive selection in PsbK was located in α-helix (Figure 7D and Figure 8C). Three amino acid sites (464th, 470th and 479th) under positive selection in RBCL were located in the random coil (Figure 7E and Figure 8D).

Meanwhile, other three genes were coding genes for protein synthesis: *rps*12 and *rps*18 genes for Ribosomal protein smaller subunit (RPS), and *rpl*22 gene for Ribosomal protein larger subunit 22 (RPL22). One positive selection site was identified in RPS12 and RPS18 protein, respectively (Figures 7G, H; Figures 8F, G). Four positive selection sites were identified in RPL22 (Figure 7F and Figure 8E). Finally, seven sites were identified in YCF1 (Hypothetical chloroplast reading frame 1) coded by *ycf*1 gene (Figure 7I and Supplementary Figure S4). Based on the protein structure prediction, most of these positive selection sites were located in the α-helix, followed by random coil and β-sheet (Figure 8).

Discussion

Chloroplast genome features

In this study, the chloroplast genomes of twelve *Paraboea* species were characterized (Figure 1; Table 2). The twelve

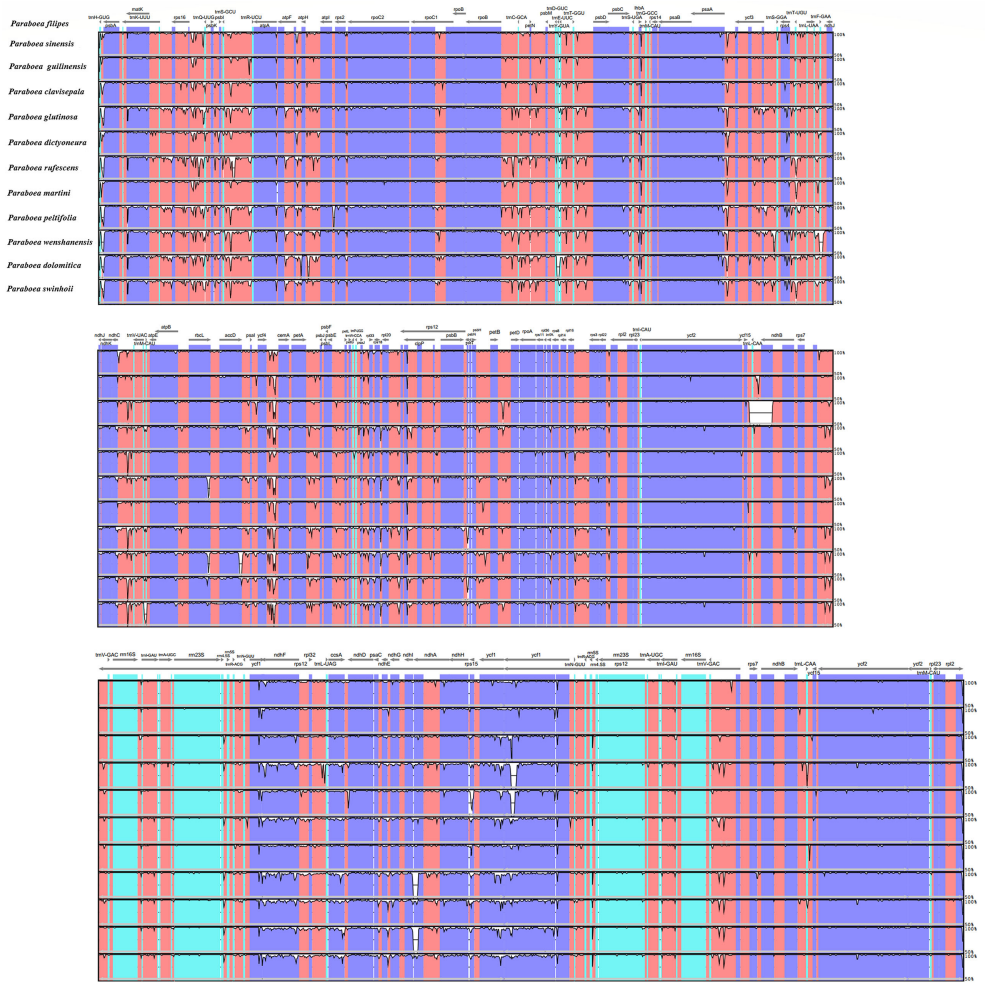


FIGURE 4
The comparative analysis with LAGAN program of the whole-chloroplast genome of twelve different species of *Paraboea*. The x-axis represents the coordinate in the chloroplast genome.

chloroplast genomes also showed a highly conserved feature in terms of structures, gene orders, gene numbers (protein-coding genes, rRNAs and tRNAs) and intron number. The chloroplast genomes of twelve *Paraboea* plants ranged from 153166 to 154245 bp in length. The chloroplast genomes of angiosperms have a highly conserved feature, but the contraction and expansion of the boundary between the IR and SC region is considered to be the main reason for the size change of the chloroplast genome (Zhang et al., 2016). The same phenomenon also existed in the twelve *Paraboea* chloroplast genomes. Despite the twelve *Paraboea* chloroplast genomes having well-conserved genomic structures including gene number and order, length variation of the whole sequences comprising IR, LSC and SSC regions was detected among these chloroplast genomes (Figure 2). In particular, *ycf1* and *ndh*F genes located at the SSC/IR border had the greatest variation in position and length in the twelve *Paraboea* chloroplast genomes. These sequence variations might be the result of boundary contraction and expansion between the SSC/IR regions in plants (Wang and Messing, 2011).

Repeat sequence analysis

SSRs have been used as molecular markers for determining a high degree of variation in similar species and are helpful to explore population genetics and polymorphisms (Zhao et al., 2015). In total, 600 SSRs were detected in the twelve chloroplast genomes, 315 of which were mononucleotide repeats, accounting for the majority of all SSRs (52.50%) (Figure 3A and Supplementary Figure S1). Among the twelve chloroplast genomes, the number of mononucleotide repeats was the largest.

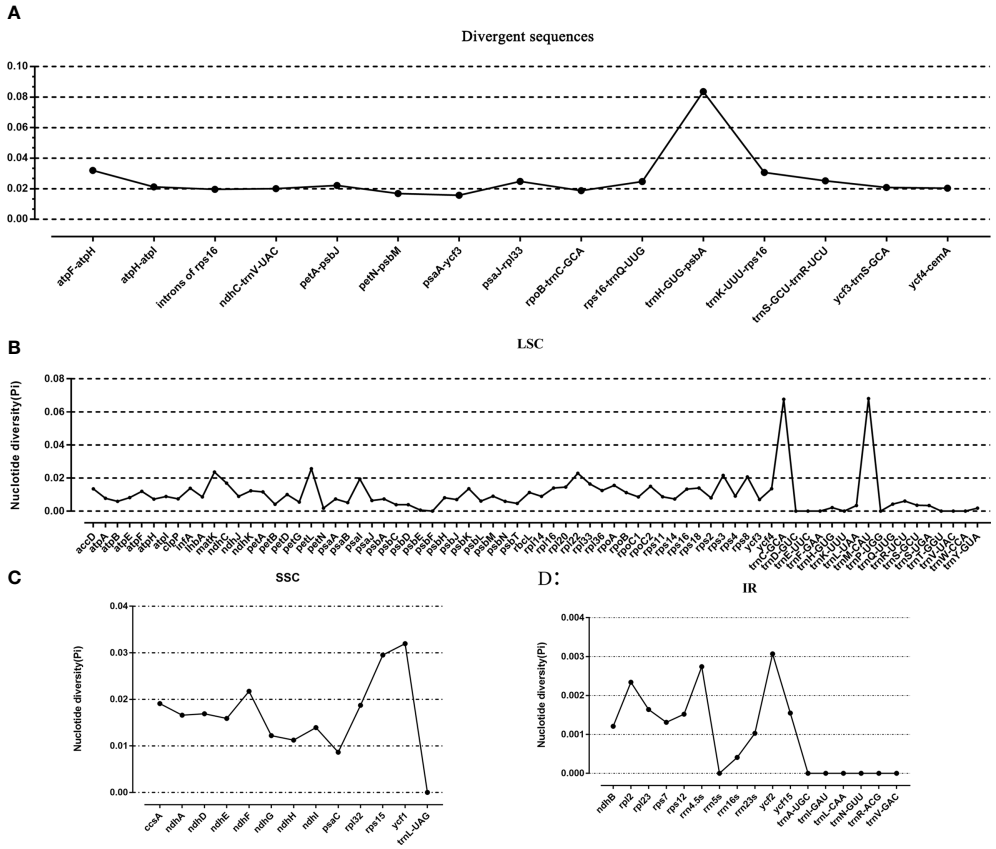


FIGURE 5
The nucleotide variability (Pi) value in the 12 *Paraboea* chloroplast genomes. **(A)** The Pi value of divergent sequences. **(B)** The Pi value of LSC region. **(C)** The Pi value of SSC region. **(D)** The Pi value of IR region.

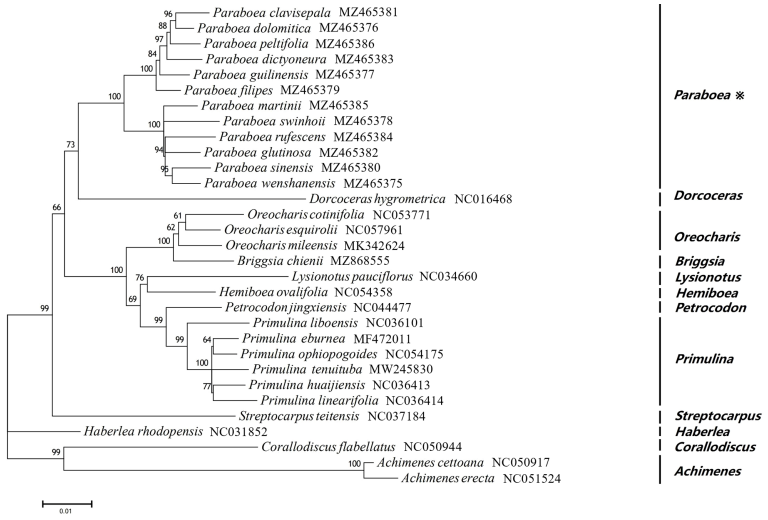


FIGURE 6
The ML phylogenetic tree of 31 Gesneriaceae species. Supporting values of > 50% for ML were shown on the branch.

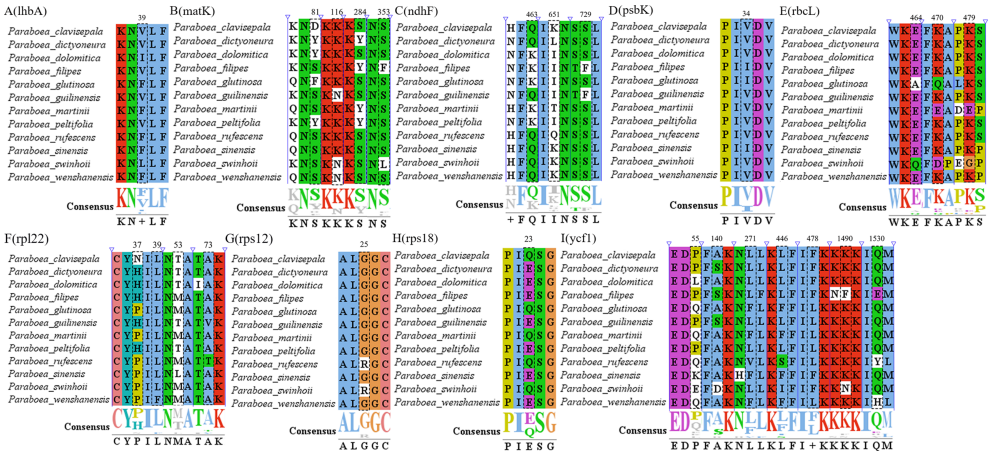


FIGURE 7
Nine genes of positive selection of amino acid sequences in site model tests.

In angiosperm chloroplast genomes, lots of similar results were also reported previously (Gandhi et al., 2010; Bessega et al., 2013). The results also demonstrated that the SSRs identified in the chloroplast genome were mostly made up of polyadenine (Poly-A) or polythymine (Poly-T) repeats, and the contents of guanine (Poly-G) and cytosine (Poly-C) repeats were low, which was consistent with the general SSR characteristics of chloroplast genomes in angiosperms (Ebert and Peakall, 2009; Asaf et al., 2020).

Moreover, 389 non-overlapped sequence repeats were identified in twelve chloroplast genomes (Figure 3B and Supplementary Figure S2), including the most non-overlapped sequence repeats (37) in *P. sinensis* and the least non-overlapped sequence repeats (28) in *P. dictyoneura*. Among the 389 repeats,

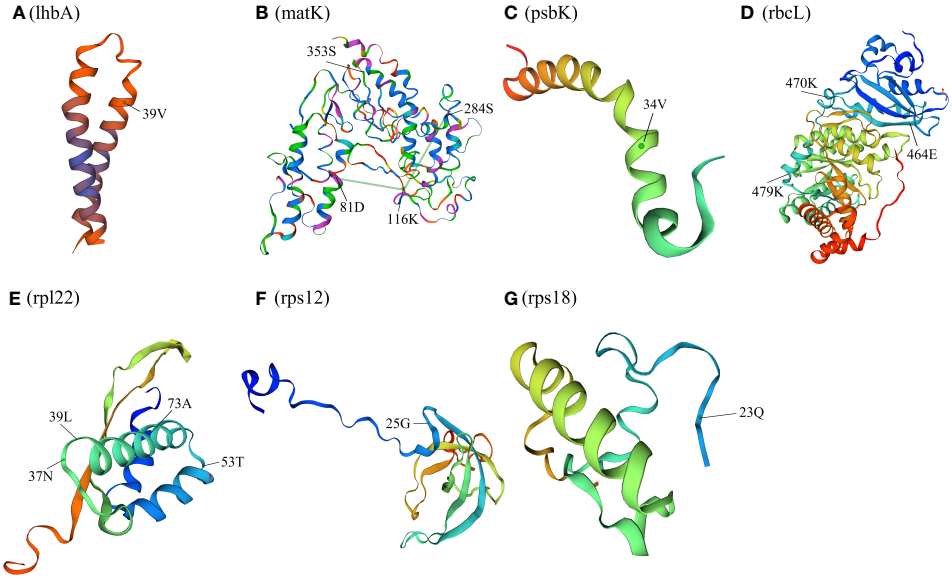


FIGURE 8
Spatial location of the positively selected sites in proteins of P. clavisepala. **(A)** Spatial location of the positively selected sites in the lhbA protein of P. clavisepala. **A, B** Spatial location of the positively selected sites in the matK protein of P. clavisepala. **A, C** Spatial location of the positively selected sites in the psbK protein of P. clavisepala. **A, D** Spatial location of the positively selected sites in the rbcL protein of P. clavisepala. **A, E** Spatial location of the positively selected sites in the rpl22 protein of P. clavisepala. **A, F** Spatial location of the positively selected sites in the rps12 protein of P. clavisepala. **A, G** Spatial location of the positively selected sites in the rps18 protein of P. clavisepala.

there were four types: forward repetition, reverse repetition, complement repetition and palindrpmic repetition. Palindrpmic repetition and forward repetition accounted for the highest proportion. The same conclusion was obtained in the analysis of repetitive sequences in the chloroplast genomes of other Gesneriaceae plants (Gu et al., 2020). All of these SSRs, together with non-overlapped sequence repeats, are useful sources to develop markers for genetic diversity analysis of *Paraboea* species (Supplementary Table S2 and Supplementary Table S3).

Phylogenetic relationship

Most of the studies on the molecular phylogeny of *Paraboea* were based on chloroplast *trn*L-F sequences and nuclear ITS sequences (Puglisi et al., 2011; Xin et al., 2019; Guo et al., 2020). Strict consensus tree based on combined ITS and *trn*L-*trn*F sequences of 53 samples showed that *Paraboea* samples formed three major clades (Puglisi et al., 2011). The major clade 1 contains Chinese and Thai *Paraboea* species, some small clades of which were with low or no branch support. The existing two chloroplast sequences didn't completely solve the phylogenetic relationship of *P. sinensis, P. rufescens, P. glanduliflora* and *P. swinhoei*.

In recent years, phylogeny based on the complete chloroplast genome has been widely used in plants (Feng et al., 2017; Kyalo et al., 2020; Tian and Wariss, 2021). 31species belonging to 12 genera of Gesneriaceae were used to construct the ML tree in this study. All nodes were supported with bootstrap values greater than 60%, and the 12 *Paraboea* species clustered into a clade with 100% bootstrap values (Figure 6), supporting the monophyly of *Paraboea*. The topology of the phylogenetic tree was more resolved than found in combined ITS and *trn*L-*trn*F datasets of previous studies (Puglisi et al., 2011; Xin et al., 2019; Guo et al., 2020). The clade formed by *P. sinensis* and *P. wenshanensis*, clustered with *P. rufescens* and *P. glutinosa*, and then shared a sister relationship with *P. martinii* and *P. swinhoei*. The sequences of chloroplast genome sequences could completely solve the phylogenetic relationship of *Paraboea*, and chloroplast genome data could provide more genetic information on the evolutionary relationships and phylogeny among species of Gesneriaceae.

Adaptive evolution analysis

The plants of *Paraboea* are mainly distributed in karst areas of south and southwest China and southeast Asia. Karst is a unique and fragile ecological environment and the rocks forming karst landforms mainly consist of limestone, dolomite and other soluble carbonate rocks. (Li et al., 2019). Because of the thin soil layer, low water holding capacity and strong permeability of rocks, the stress of frequent alternation of dry and wet is common in karst habitats. Facing frequent temporary drought, karst plants generally appear enhanced photosynthetic capacity and light protection mechanisms (Liu et al., 2021). The changes in ground temperature, air temperature, light intensity and atmospheric relative humidity are quite different in different microhabitats (such as rocky hills with direct sunlight, forests with weak light and dark caves, etc.) (Ou et al., 2020). These challenging karst environments may impose selective pressure on genes, which could leave a footprint of natural selection in genes of chloroplast involved in adaptation to the environment. In this study, among the chloroplast genes of twelve *Paraboea* species, nine genes (*lhb*A, *mat*K, *ndh*F, *psb*K, *rbc*L, *rpl22*, *rps12*, *rps18* and *ycf1*) were identified under positive selection using a site model (Figure 7, Figure 8 and Supplementary Table S4).

The maturase encoded by *mat*K gene is involved in splicing of introns of *trn*K, *trn*I, *atp*F and other genes, which is important for maintaining the normal function of chloroplast (Moran et al., 1994; Vogel et al., 1999; Lambowitz and Zimmerly, 2004; Stern et al., 2010; Zoschke et al., 2010). There were four positive selection sites in *mat*K gene of *Paraboea* species, and *mat*K gene also had undergone adaptive evolution in Lycopodiaceae, Bryophyta and other plants (Hao et al., 2010a; Hao et al., 2010b). Adaptive evolution of *mat*K may fine-tune its function to optimize its performance in various environmental conditions.

Three genes under positive selection were related to photosynthesis, namely *psb*K, *ndh*F and *rbc*L gene. The *psb*K gene encodes Photosystem II subunits K. Photosystem II is the first link in the chain of photosynthesis, and captures photons and uses the energy to extract electrons from water molecules (Ferreira et al., 2004). PSBK is not necessary for the assembly or activity of photosystem II complex, but is essential for optimal photosystem II function. The *psb*K gene was detected under positive selection in *Echinacanthus* (Acanthaceae) and *Calligonum Mongolicum* (Polygonaceae), and speculated to play an important role in plant adaptation evolutionary process to the diverse environment (Gao et al., 2019a; Duan et al., 2020). The *ndh*F gene encodes NADH dehydrogenase subunit protein (Kubicki et al., 1996). In previous studies on plant adaptive evolution, *ndh*F genes were often under positive selection pressure (Liu et al., 2020; Li et al., 2021; Wen et al., 2021). The NADH dehydrogenase complex of higher plants not only participated in photosynthetic electron transport (Joet et al., 2001; Joet, 2002), but also acted as an electron transport carrier for chloroplast respiration (Casano et al., 2000). The adaptive evolution of the *ndh*F gene may affect energy transformation and resistance to photooxidative stress in different environments. The *rbc*L gene encoded the gene coding for the rubisco large subunit protein of Rubisco, which was an important part of the photosynthesis electron transport regulator (Piot et al., 2018). The *rbc*L gene was often under positive selection because of being the target of selection diverse environment factors related to the changes in temperature, drought and carbon dioxide concentration (Fan et al., 2018). NADH-dehydrogenase subunits

and Photosystem subunits were essential in the electron transport chain for the generation of ATP and light energy utilization, which were all indispensable parts for photosynthesis of plants (Yamori and Shikanai, 2016; Peltier et al., 2016). Therefore, the signature of positive selection in three genes related to photosynthesis suggests that they might have been involved in adaptation to diversified environments for *Paraboea* species in karst habitats.

Meanwhile, positive selection sites were also identified in *lhb*A, *rpl22*, *rps12* and *rps18* genes. The specific function of *lhb*A gene has not been fully studied (Wu et al., 2020). The *rps* genes encode small ribosomal subunit proteins, and *rpl* genes encode large ribosomal subunit proteins (Muto and Ushida, 1995). The mutation of genes encoded in ribosomal proteins under the pressure of the natural environment may affect the translation of chloroplast ribosome (Ramundo et al., 2013).

Seven sites were detected under positive selection in the *ycf1* gene. Positive selection of *ycf1* was also found to be involved in the adaptation of the genus *Panax* (Jiang et al., 2018). Being one of the largest chloroplast genes, the *ycf1* gene encoding a component of the chloroplast's inner envelope membrane protein translocon, has become a useful gene for assessing sequence variations and evolutionary processes in plants (Huang et al., 2010; Kikuchi et al., 2013). The function and the adaptive evolutionary analysis of the *ycf1* gene would better understand the evolutionary mechanism of plants in the future.

Because of environmental pressure, adaptive evolution of chloroplast genomes is a common phenomenon, especially for genes involved in photosynthesis. Genes associated with photosynthesis are more likely to evolve adaptively in plants distributed in extreme environments, such as shade plants or aquatic plants (Xie et al., 2018). In karst areas, there are great differences in environmental factors such as light intensity, soil water content and nutrient availability, which might have exerted strong selective forces on plant evolution (Ai et al., 2015). In this study, nine chloroplast genes under positive selection, most of which were related to photosynthesis and protein synthesis, may possibly contribute to the diverse evolution and adaptation of *Paraboea* species to karst extreme environments.

Conclusion

This is the first report of the complete chloroplast genome sequence of *Paraboea* species. In this study, the newly sequenced chloroplast genomes of twelve *Paraboea* species were reported and compared. The genome annotation and comparative analysis showed that each chloroplast genome was a typical quadripartite structure like traditional angiosperms, and the GC content, gene number and order were similar to each other. The chloroplast genomes of the twelve *Paraboea* species were similar in structure, composition and gene order. In the twelve *Paraboea* chloroplast genomes, a total of 600 SSRs and 389 non-overlapped sequence repeats were identified, which were informative sources for developing markers for genetic diversity analysis of *Paraboea* species. In addition, we found that 10 different regions (*trn*H-GUG-*psb*A, *trn*M-CAU, *trn*C-GCA, *atp*F-*atp*H, *ycf1*, *trn*K-UUU-*rps16*, *rps15*, *pet*L, *trn*S-GCU-*trn*R-UCU and *psa*J-*rpl33*) were potential molecular markers in twelve *Paraboea* species. The phylogenetic tree based on 76 protein coding genes clearly demonstrated the genetic and evolutionary relationships of 31 species belonging to 12 genera of Gesneriaceae. Adaptive evolution analysis detected positive selection signals in nine chloroplast genes (i.e., *lhb*A, *mat*K, *ndh*F, *psb*K, *rbc*L, *rpl22*, *rps12*, *rps18* and *ycf1*). The evolution of *Paraboea* to adapt to extreme habitats in karst environments may be linked to changes in these positive selection sites. These analyses of chloroplast genomes will provide preparations for the development and utilization of *Paraboea* species germplasm resources and the formulation of conservation strategies.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Author contributions

BZ, FW, YW, YM and XH conceived and designed the study. ZL, YW collected and analyzed the data. BZ, FW, YW, ZL and XH wrote the manuscript. All authors have directly contributed to this manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study is supported by the Key Sci. & Tech. Research and Development Project of Guangxi (Guike AD20159091 &

ZY21195050), the Guangxi Natural Science Foundation (2020GXNSFBA297049), the capacity-building project of SBR of CAS (KFJ-BRP-017-68), the Anhui Provincial Natural Science Foundation (1908085QC1), the Fund of Yunnan Key Laboratory for Integrative Conservation of Plant Species with Extremely Small Populations (PSESP2021F07).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1019831/full#supplementary-material

SUPPLEMENTARY FIGURE 1
The comparison of each type of SSRs among 12 *Paraboea* choloplast genomes.

SUPPLEMENTARY FIGURE 2
Number of the four complex repeat types (forward, palindrome, reverse, and complement) in the twelve *Paraboea* chloroplast.

SUPPLEMENTARY FIGURE 3
Protein secondary structure of *ndh*F.

SUPPLEMENTARY FIGURE 4
Protein secondary structure of *ycf1*.

SUPPLEMENTARY TABLE 1
Chloroplast genome sequences from GenBank used in this study.

SUPPLEMENTARY TABLE 2
Summary of SSRs in twelve *Paraboea* chloroplast genomes.

SUPPLEMENTARY TABLE 3
Summary of complex repeats in twelve *Paraboea* chloroplast genomes.

SUPPLEMENTARY TABLE 4
Positive selection sites identified in the chloroplast genomes of twelve *Paraboea* species.

References

Ai, B., Gao, Y., Zhang, X., Tao, J., Kang, M., and Huang, H. (2015). Comparative transcriptome resources of eleven *Primulina* species, a group of 'stone plants' from a biodiversity hot spot. *Mol. Ecol. Resour.* 15, 619–632. doi: 10.1111/1755-0998.12333

Amiryousefi, A., Hyvönen, J., and Poczai, P. (2018). IRscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics* 34, 3030–3031. doi: 10.1093/bioinformatics/bty220

Asaf, S., Jan, R., Khan, A. L., and Lee, I. J. (2020). Complete chloroplast genome characterization of *Oxalis corniculata* and its comparison with related species from family oxalidaceae. *Plants* 9, 928. doi: 10.3390/plants9080928

Beier, S., Thiel, T., Munch, T., Scholz, U., and Mascher, M. (2017). MISA-web: a web server for microsatellite prediction. *Bioinformatics* 33 (16), 2583–2585. doi: 10.1093/bioinformatics/btx198

Bessega, C. F., Pometti, C. L., Miller, J. T., Watts, R., Saidman, B. O., and Vilardi, J. C. (2013). New microsatellite loci for *Prosopis alba* and *P. chilensis* (Fabaceae). *Appl. Plant Sci.* 1 (5), 1200324. doi: 10.3732/apps.1200324

Buchan, W. A., Minneci, F., Nugent, C. O., Bryson, K., and Jones, D. T. (2013). Scalable web services for the PSIPRED protein analysis workbench. *Nucleic Acids Res.* 41, W349–W357. doi: 10.1093/nar/gkt381

Casano, L. M., Zapata, J. M., Martin, M., and Sabater, B. (2000). Chlororespiration and poising of cyclic electron transport. plastoquinone as electron transporter between thylakoid NADH dehydrogenase and peroxidase. *J. Biol. Chem.* 275 (2), 942–948. doi: 10.1074/jbc.275.2.942

Chen, H. M., Chen, Z. E., Du, Q., Jiang, M., Wang, B., and Liu, C. (2022). Complete chloroplast genome of *Campsis grandiflora* (Thunb.) schum and systematic and comparative analysis within the family bignoniaceae. *Mol. Biol. Rep.* 49 (4), 3085–3098. doi: 10.1007/s11033-022-07139-0

Chen, J., Zang, Y., Shang, S., Liang, S., Zhu, M., Wang, Y., et al. (2021). Comparative chloroplast genomes of *Zosteraceae* species provide adaptive evolution insights into seagrass. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.741152

Dierckxsens, N., Mardulyn, P., and Smits, G. (2017). NOVOPlasty: *de novo* assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* 45 (4), e18. doi: 10.1093/nar/gkw955

Doyle, J. J., and Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* 19, 11–15.

Duan, H. R., Zhang, Q., Yang, H., Tian, F. P., Hu, Y., Wang, C. M., et al. (2020). Complete chloroplast genome of *Calligonum mongolicum*: Genome organization, codon usage pattern, phylogenetic relationships, comparative structure and adaptive evolution analysis. *Research Square*. doi: 10.21203/rs.3.rs-49271/v1

Ebert, D., and Peakall, R. (2009). Chloroplast simple sequence repeats (cpSSRs): Technical resources and recommendations for expanding cpSSR discovery and applications to a wide array of plant species. *Mol. Ecol. Res.* 9, 673–690. doi: 10.1111/j.1755-0998.2008.02319.x

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32 (5), 1792–1797. doi: 10.1093/nar/gkh340

Fan, W. B., Wu, Y., Yang, J., Shahzad, K., and Li, Z. H. (2018). Comparative chloroplast genomics of *Dipsacales* species: Insights into sequence variation, adaptive evolution and phylogenetic relationships. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00689

Feng, C., Xu, M., Feng, C., and Kang, M. (2017). The complete chloroplast genome of *Primulina* and two novel strategies for development of high polymorphic loci for population genetic and phylogenetic studies. *BMC evol. Biol.* 17 (1). doi: 10.3389/fpls.2018.00689

Feng, S. G., Zheng, K. X., Jiao, K. L., Cai, Y. C., Chen, C. L., Mao, Y. Y., et al. (2020). Complete chloroplast genomes of four *Physalis* species (Solanaceae): lights into genome structure, comparative analysis, and phylogenetic relationships. *BMC Plant Biol.* 20 (1), 242. doi: 10.1186/s12870-020-02429-w

Ferreira, K. N., Iverson, T. M., Maghlaoui, K., Barber, J., and Iwata, S. (2004). Architecture of the photosynthetic oxygen-evolving center. *Science* 303 (5665), 1831–1838. doi: 10.1126/science.1093087

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., and Dubchak, I. (2004). VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* 32, W273–W279. doi: 10.1093/nar/gkh458

Gandhi, S. G., Awasthi, P., and Bedi, Y. S. (2010). Analysis of SSR dynamics in chloroplast genomes of brassicaceae family. *Bioinformation* 5 (1), 16–20. doi: 10.6026/97320630005016

Gao, F., Chen, C., Arab, D., Du, Z., He, Y., and Ho, S. (2019b). EasyCodeML: a visual tool for analysis of selection using CodeML. *Ecol. Evol.* 9, 3891–3898. doi: 10.1002/ece3.5015

Gao, C., Deng, Y. , and Wang, J. (2019a). The complete chloroplast genomes of *Echinacanthus* species (Acanthaceae): Phylogenetic relationships, adaptive evolution, and screening of molecular markers. *Front. Plant Sci.* 91989. doi: 10.3389/fpls.2018.01989

Gao, J. Y., Ren, P. Y., Yang, Z. H., and Li, Q. J. (2006). The pollination ecology of *Paraboea rufescens* (Gesneriaceae): a buzz-pollinated tropical herb with mirror-image flowers. *Ann. botany* 97 (3), 371–376. doi: 10.1093/aob/mcj044

Greiner, S., Lehwark, P., and Bock, R. (2019). OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* 47 (W1), W59–W64. doi: 10.1093/nar/gkz238

Gu, C. H., Dong, B., Xu, L., Tembrock, L. R., Zheng, S. Y., and Wu, Z. Q. (2018). The complete chloroplast genome of *H. myrtifolia* and comparative analysis within myrtales. *Molecules1019831* 23, 846. doi: 10.3390/molecules23040846

Guo, J., Meng, T., Pang, H., and Zhang, Q. (2016). *Petrocodon retroflexus* sp. nov. (gesneriaceae) from a karst cave in guizhou, China. *Nordic J. Botany* 34 (2), 159–164. doi: 10.1111/njb.00941

Guo, Z., Wu, Z., Xu, W., Li, Z., and Xiang, X. (2020). *Paraboea dolomitica* (Gesneriaceae), a new species from guizhou, China. *PhytoKeys* 153, 37–48. doi: 10.3897/phytokeys.153.50933

Gu, L., Su, T., An, M. T., and Hu, G. X. (2020). The complete chloroplast genome of the vulnerable *Oreocharis esquirolii* (Gesneriaceae): Structural features, comparative and phylogenetic analysis. *Plants (Basel)* 9 (12), 1692. doi: 10.3390/plants9121692

Hao, D. C., Chen, S. L. , and Xiao, P. G. (2010a). Molecular evolution and positive Darwinian selection of the chloroplast maturase *mat*K. *J. Plant Res.* 123 (2), 241–247. doi: 10.1007/s10265-009-0261-5

Hao, D. C., Mu, J., Chen, S. L. , and Xiao, P. G. (2010b). Physicochemical evolution and positive selection of the gymnosperm *mat*K proteins. *J. Genet.* 89 (1), 81–89. doi: 10.1007/s12041-010-0014-1

Huang, J. L., Sun GL,, and Zhang, D. M. (2010). Molecular evolution and phylogeny of the angiosperm *ycf2* gene. *J. Syst. Evol.* 48, 240–248. doi: 10.1111/j.1759-6831.2010.00080.x

Huelsenbeck, J., and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17, 754–755. doi: 10.1093/bioinformatics/17.8.754

Jiang, P., Shi, F. X., Li, M. R., Liu, B., Wen, J., Xiao, H. X., et al. (2018). Positive selection driving cytoplasmic genome evolution of the medicinally important ginseng plant genus *Panax*. *Front. Plant sci.* 9. doi: 10.3389/fpls.2018.00359

Joet, T. (2002). Cyclic electron flow around photosystem I in C3 plants. *In vivo* control by the redox state of chloroplasts and involvement of the NADH-dehydrogenase complex. *Plant Physiol.* 128 (2), 760–769. doi: 10.1104/pp.010775

Joet, T., CournAc, L., Horvath, E. M., and Peltier, M. G. (2001). Increased sensitivity of photosynthesis to antimycin induced by inactivation of the chloroplast ndhB gene. evidence for a participation of the NADH-dehydrogenase complex to cyclic electron flow around photosystem I. *Plant Physiol.* 125 (4), 1919–1929. doi: 10.1104/pp.125.4.1919

Kazutaka, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kikuchi, S., Bédard, J., Hirano, M., Hirabayashi, Y., Oishi, M., Imai, M., et al. (2013). Uncovering the protein translocon at the chloroplast inner envelope membrane. *Science* 339, 571–574. doi: 10.1126/science.1229262

Kubicki, A., Funk, E., and Steinmüller, W. K. (1996). Differential expression of plastome-encoded ndh genes in mesophyll and bundle-sheath chloroplasts of the C4plant sorghum bicolor indicates that the complex I-homologous NAD(P)H-plastoquinone oxidoreductase is involved in cyclic electron transport. *Planta* 199 (2), 276–281. doi: 10.1002/anie.200905829

Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* 29, 4633–4642. doi: 10.1093/nar/29.22.4633

Kyalo, C. M., Li, Z. Z., Mkala, E. M., Malombe, I., Hu, G. W., and Wang, Q. F. (2020). The first glimpse of *Streptocarpus ionanthus* (Gesneriaceae) phylogenomics: Analysis of five subspecies' chloroplast genomes. *Plants* 9 (4), 456. doi: 10.3390/plants9040456

Lambowitz, A. M., and Zimmerly, S. (2004). Mobile group II introns. *Annu. Rev. Genet.* 38, 1–35. doi: 10.1146/annurev.genet.38.072902.091600

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10 (3), R25. doi: 10.1186/gb-2009-10-3-r25

Li, F., He, X., Sun, Y., Zhang, X., Tang, X. , Li, Y., et al. (2019). Distinct endophytes are used by diverse plants for adaptation to karst regions. *Sci. Rep.* 9 (1), 5246. doi: 10.1038/s41598-019-41802-0

Li, J. L., Tang, J. M., Zeng, S. Y., Han, F., Yuan, J., and Yu, J. (2021). Comparative plastid genomics of four *Pilea* (Urticaceae) species: insight into interspecific plastid genome diversity in pilea. *BMC Plant Biol.* 21 (1), 25. doi: 10.1186/s12870-020-02793-7

Liu, C., Huang, Y., Wu, F., Liu, W., Ning, Y., Huang, Z., et al. (2021). Plant adaptability in karst regions. *J. Plant Res.* 134 (5), 889–906. doi: 10.1007/s10265-021-01330-3

Liu, Q., Li, X., Li, M., Xu, W., and Heslop-Harrison, J. S. (2020). Comparative chloroplast genome analyses of *Avena*: insights into evolutionary dynamics and phylogeny. *BMC Plant Biol.* 20 (1), 406. doi: 10.1186/s12870-020-02621-y

Li, J. M., and Wang, Y. Z. (2007). Phylogenetic reconstruction among species of *Chiritopsis* and *Chirita sect. gibbosaccus* (Gesneriaceae) based on nrDNA ITS and cpDNA trnL-f sequences. *Syst. Botany* 32 (4), 888–898. doi: 10.1600/036364407783390764

Moran, J. V., Mecklenburg, K. L., Sass, P., Belcher, S. M., Mahnke, D., Lewin, A., et al. (1994). Splicing defective mutants of the COXI gene of yeast mitochondrial DNA: initial definition of the maturase domain of the group II intron AI2. *Nucleic Acids Res.* 22 (11), 2057–2064. doi: 10.1093/nar/22.11.2057

Muto, A., and Ushida, C. (1995). Transcription and translation. *Methods Cell Biol.* 48, 483. doi: 10.1186/s12870-020-02621-y

Ou, Z., Pang, S., He, Q., Peng, Y., Huang, X., and Shen, W. (2020). Effects of vegetation restoration and environmental factors on understory vascular plants in a typical karst ecosystem in southern China. *Sci. Rep.* 10 (1), 1–10. doi: 10.1038/s41598-020-68785-7

Peltier, G., Aro, E. M., and Shikanai, T. (2016). NDH-1 and NDH-2 plastoquinone reductases in oxygenic photosynthesis. *Annu. Rev. Plant Biol.* 67, 55–80. doi: 10.1146/annurev-arplant-043014-114752

Piot, A., Hackel, J., Christin, P. A., and Besnard, G. (2018). One-third of the plastid genes evolved under positive selection in PACMAD grasses. *Planta* 247, 255–266. doi: 10.1007/s00425-017-2781-x

Puglisi, C., Middleton, D. J., Triboun, P., and Möller, M. (2011). New insights into the relationships between *Paraboea*, *Trisepalum* and *Phylloboea* (Gesneriaceae) and their taxonomic consequences. *Taxon* 60 (6), 1693–1702. doi: 10.1002/tax.606014

Puglisi, C., and Phutthai, T. (2017). A new species of *Paraboea* (Gesneriaceae) from Thailand. *Edinburgh J. Botany* 75 (1), 51–54. doi: 10.1017/S0960428617000324

Ramundo, S., Rahire, M., Schaad, O., and Rochaix, J. D. (2013). Repression of essential chloroplast genes reveals new signaling pathways and regulatory feedback loops in *Chlamydomonas*. *Plant Cell.* 25, 167–186. doi: 10.1105/tpc.112.103051

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., et al. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol. Biol. Evol.* 34 (12), 3299–3302. doi: 10.1093/molbev/msx248

Schattner, P., Brooks, A. N., and Lowe, T. M. (2005). The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* 33, W686–W689. doi: 10.1093/nar/gki366

Shi, L., Chen, H., Jiang, M., Wang, L., Wu, X., Huang, L., et al. (2019). CPGAVAS2, an integrated plastome sequence annotator and analyzer. *Nucleic Acids Res.* 47 (W1), W65–w73. doi: 10.1093/nar/gkz345

Song, Y., Zhao, W. J., Xu, J., Li, M. F., and Zhang, Y. J. (2022). Chloroplast genome evolution and species identification of *Styrax* (Styracaceae). *BioMed. Res. Int.* 22, 5364094. doi: 10.1155/2022/5364094

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22, 2688–2690. doi: 10.1093/bioinformatics/btl446

Stern, D. B., Goldschmidt-Clermont, M., and Hanson, M. R. (2010). Chloroplast RNA metabolism. *Annu. Rev. Plant Biol.* 61 (1), 125–155. doi: 10.1146/annurev-arplant-042809-112242

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst197

Tao, J., Feng, C., Ai, B., and Kang, M. (2016). Adaptive molecular evolution of the two-pore channel 1 gene TPC1 in the karst-adapted genus *Primulina* (Gesneriaceae). *Ann. botany* 118 (7), 1257–1268. doi: 10.1093/aob/mcw168

Tian, X., and Wariss, H. M. (2021). The complete chloroplast genome sequence of *Metabriggsia ovalifolia* w. t. Wang (Gesneriaceae), a national key protected plant

endemic to karst areas in China. *Mitochondrial DNA B Resour.* 6 (3), 833–834. doi: 10.1080/23802359.2021.1884021

Vogel, J., Borner, T., and Hess, W. R. (1999). Comparative analysis of splicing of the complete set of chloroplast group II introns in three higher plant mutants. *Nucleic Acids Res.* 27 (19), 3866–3874. doi: 10.1093/nar/27.19.3866

Wang, W., and Messing, J. (2011). High-throughput sequencing of three lemnoideae (duckweeds) chloroplast genomes from. *PloS One* 6, e24670. doi: 10.1371/journal.pone.0024670

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., et al. (2018). SWISS- MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46, W296–W303. doi: 10.1093/nar/gky427

Wen, F., Hong, X., Chen, L. Y., Zhou, S. B., and Wei, Y. G. (2013). A new species of *Paraboea* (Gesneriaceae) from a karst limestone hill in southwestern guangdong, China. *Phytotaxa* 131 (1), 1–8. doi: 10.11646/phytotaxa.131.1.1

Wen, F., Wu, X., Li, T., Jia, M., Liu, X., and Liao, L. (2021). The complete chloroplast genome of *Stauntonia chinensis* and compared analysis revealed adaptive evolution of subfamily lardizabaloideae species in China. *BMC Genomics* 22 (1), 161. doi: 10.1186/s12864-021-07484-7

Wu, Z. H., Liao, R., Yang, T. G., Dong, X., Lan, D. Q., Qin, R., et al. (2020). Analysis of six chloroplast genomes provides insight into the evolution of *Chrysosplenium* (Saxifragaceae). *BMC Genomics* 21 (1), 621. doi: 10.1186/s12864-020-07045-4

Xie, D. F., Yu, Y., Deng, Y. Q., Li, J., Liu, H. Y., Zhou, S. D., et al. (2018). Comparative analysis of the chloroplast genomes of the Chinese endemic genus *Urophysa* and their contribution to chloroplast phylogeny and adaptive evolution. *Int. J. Mol. Sci.* 19 (7), 1847. doi: 10.3390/ijms19071847

Xin, Z. B., Chou, W. C., Maciejewski, S., Fu, L. F., and Wen, F. (2021). *Primulina papillosa* (Gesneriaceae), a new species from limestone areas of guangxi, China. *PhytoKeys* 177, 55–61. doi: 10.3897/arphapreprints.e63933

Xin, Z. B., Fu, L. F., Fu, Z. X., Li, S., Wei, Y. G., Wen, F., et al. (2019). Complete chloroplast genome sequence of *Petrocodon jingxiensis* (Gesneriaceae). *Mitochondrial DNA Part B.* 4 (2), 2771–2772. doi: 10.1080/23802359.2019.1624208

Xu, W. B., Huang, Y. S., Wei, G. F., Tan, W. N., and Liu, Y. (2012). *Paraboea angustifolia* (Gesneriaceae): A new species from limestone areas in northern guangxi, China. *Phytotaxa* 62 (1), 39–43. doi: 10.1007/s12228-010-9175-8

Yamori, W., and Shikanai, T. (2016). Physiological functions of cyclic electron transport around photosystem I in sustaining photosynthesis and plant growth. *Annu. Rev. Plant Biol.* 67, 81–106. doi: 10.1146/annurev-arplant-043015-112002

Yang, X., Xie, D. F., Chen, J. P., Zhou, S. D., Yu, Y., and He, X. J. (2020). Comparative analysis of the complete chloroplast genomes in *Allium* subgenus *Cyathophora* (Amaryllidaceae): Phylogenetic relationship and adaptive evolution. *BioMed. Res. Int.* 20, 1732586. doi: 10.1155/2020/1732586

Zhai, Y. F., Yu, X. Q., Zhou, J. G., Li, J., Tian, Z., Wang, P. Q., et al. (2021). Complete chloroplast genome sequencing and comparative analysis reveals changes to the chloroplast genome after allopolyploidization in *Cucumis*. *Genome* 64 (6), 627–638. doi: 10.1139/gen-2020-0134

Zhang, Y. J., Du, L. W., Liu, A., Chen, J. J., Wu, L., Hu, W. M., et al. (2016). The complete chloroplast genome sequences of five *Epimedium* species: lights into phylogenetic and taxonomic analyses. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.00306

Zhang, R., Zhang, L., Wang, W., Zhang, Z., Du, H., Qu, Z., et al. (2018). Differences in codon usage bias between photosynthesis-related genes and genetic system-related genes of chloroplast genomes in cultivated and wild *Solanum* species. *Int. J. Mol. Sci.* 19 (10), 3142. doi: 10.3390/ijms19103142

Zhao, Y. B., Yin, J. L., Guo, H. Y., Zhang, Y. Y., Xiao, W., Sun, C., et al. (2015). The complete chloroplast genome provides insight into the evolution and polymorphism of *Panax ginseng*. *Front. Plant Sci.* 5. doi: 10.3389/fpls.2014.00696

Zhou, P., Gu, Z. J., and Möller, M. (2003). New chromosome counts and nuclear characteristics for some members of gesneriaceae subfamily cyrtandroideae from China and Vietnam. *Edinburgh J. Botany* 60 (3), 449–466. doi: 10.1017/S0960428603000349

Zoschke, R., Nakamura, M., Liere, K., Sugiura, M., Börner, T., Schmitz-Linneweber, C., et al. (2010). An organellar maturase associates with multiple group II introns. *Proc. Natl. Acad. Sci. U. S. A.* 107 (7), 3245–3250. doi: 10.1073/pnas.0909400107

frontiers | Frontiers in Plant Science

TYPE Technology and Code
PUBLISHED 03 November 2022
DOI 10.3389/fpls.2022.1011948

OPEN ACCESS

EDITED BY
Gang Zhang,
Shaanxi University of Chinese Medicine, China

REVIEWED BY
Rosario Carmona,
Andalusian Public Foundation Progress and Health Hospital Universitario Virgen del Rocío, Spain
Abdullah.,
Quaid-i-Azam University, Pakistan
Yongqi Zheng,
Chinese Academy of Forestry, China

*CORRESPONDENCE
Martina Strömvik
martina.stromvik@mcgill.ca

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

RECEIVED 04 August 2022
ACCEPTED 14 October 2022
PUBLISHED 03 November 2022

CITATION
Chen W, Achakkagari SR and Strömvik M (2022) Plastaumatic: Automating plastome assembly and annotation.
Front. Plant Sci. 13:1011948.
doi: 10.3389/fpls.2022.1011948

Plastaumatic: Automating plastome assembly and annotation

Wenyi Chen†, Sai Reddy Achakkagari† and Martina Strömvik*

Department of Plant Science, McGill University, Sainte-Anne-de-Bellevue, QC, Canada

Plastome sequence data is most often extracted from plant whole genome sequencing data and need to be assembled and annotated separately from the nuclear genome sequence. In projects comprising multiple genomes, it is labour intense to individually process the plastomes as it requires many steps and software. This study developed *Plastaumatic* - an automated pipeline for both assembly and annotation of plastomes, with the scope of the researcher being able to load whole genome sequence data with minimal manual input, and therefore a faster runtime. The main structure of the current automated pipeline includes trimming of adaptor and low-quality sequences using *fastp*, *de novo* plastome assembly using *NOVOPlasty*, standardization and quality checking of the assembled genomes through a custom script utilizing *BLAST+* and *SAMtools*, annotation of the assembled genomes using *AnnoPlast*, and finally generating the required files for NCBI GenBank submissions. The pipeline is demonstrated with 12 potato accessions and three soybean accessions.

KEYWORDS

plastome, organellar genome, chloroplast genome, sequence assembly, plant

Introduction

Plastids are essential organelles in plant cells as they host the vital reactions of photosynthesis (as chloroplasts), store starch and sugars (amyloplasts), lipids and oils (elaioplasts), as well as pigments (chromoplasts). All differentiated plastid types develop from the proplastid. Just like the mitochondrion, the (pro)plastid has its own genome, also known as the plastome. The plastome of most land plants is relatively conserved in size and structure – a circular molecule in the size range of 120,000 to 170,000 base pairs. It usually consists of four structural regions including one large single copy (LSC), one small single copy (SSC), and two inverted repeat regions (IRa and IRb) (Chung et al., 2006). Being highly conserved across species, the genetic information contained in the plastome could hold keys to a better understanding of plant adaptation, as well as crop

improvement and breeding. Being generally inherited maternally (just like the mitogenome), the plastome is often extensively studied in phylogenetic analyses of plants (McCauley et al., 2007).

Plastome assembly typically includes the following manually initiated steps: the trimming of adaptors and low-quality sequences from whole genome sequencing data using tools such as *Trimmomatic* (Bolger et al., 2014), *de novo* plastome assembly using the most popular tool *NOVOPlasty* (Dierckxsens et al., 2017), or GetOrganelle (Jin et al., 2020) and annotation of the assembled genomes with well-annotated reference plastomes using *PGA* (Qu et al., 2019) or *GeSeq* (Tillich et al., 2017), where the running of each tool mentioned above requires a written script specifying paths of input and output files, the executing commands and modified parameters.

As more and more projects sequence multiple plant genomes for comparison and need to assemble the corresponding plastomes (Achakkagari et al., 2020; Achakkagari et al., 2021; Camargo Tavares et al., 2022; Hoopes et al., 2022), the time spent on tedious and repeated manual input and sorting can be avoided if the process was automated. An automated workflow for fast and accurate assembly as well as annotation of plastome sequences from raw whole (nuclear) genome sequencing data is needed.

Currently there are no automated pipelines for the assembly and annotation of the plastomes. For example, the pipeline *NOVOWrap*, (Wu et al., 2021a) is available publicly and can assemble and standardize the plastome sequences, however it does not incorporate trimming and annotation methods in the pipeline. The Fast-Plast (McKain and Wilson, 2017) is another similar tool, which however also does not incorporate annotation in its pipeline.

In the current study an automated pipeline for both assembly and annotation of plastomes was developed, with the scope of the researcher being able to load whole (nuclear) genome sequence data from any number of genotypes, species, or related organisms at a time, with minimal manual input, and therefore a faster completion rate. The pipeline is demonstrated with two sets of plant sequence data: three soybean accessions, and 12 potato accessions, and shows substantially faster completion than manual assembly.

Methods

The automation of the pipeline was achieved through Snakemake, a specification language built on Python (Mölder et al., 2021). A snakefile outlining rules that describe steps in a workflow defining how to obtain output files from input files. Dependencies between rules are determined automatically according to the manner the snakefile was written. Upon executing the snakefile, Snakemake can then run through all described steps in the workflow at once by taking the output files from an upstream rule and automatically feed them into the next rule. This automated pipeline for plastome assembly and annotation was made automatic through specifying the connections of input and output files for each program to those of the next and previous program. The automated processes in this pipeline specified in the main executable snakefile include six steps: quality trimming by *fastp*, generating input config files for *de novo* assembly, *de novo* assembly by *NOVOPlasty*, standardization of the assembled genomes using a custom script, annotation of the assembled plastomes by *AnnoPlast.py*, and GenBank to feature table conversion using *gbf2tbl.pl* script from NCBI tools (Figure 1).

User input

In order to execute the pipeline, the computing systems used need to have Snakemake installed (https://snakemake.readthedocs.io/en/stable/getting_started/installation.html). Executing the pipeline requires a configuration file from the user. The configuration file requests the paths of the forward and reverse raw reads in *fastq* format (compressed or uncompressed), a seed sequence in *fasta* format from a closely related reference plastome

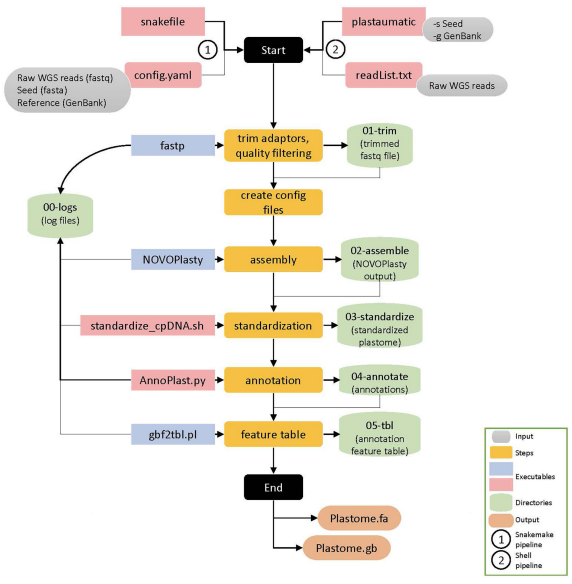


FIGURE 1
Schematic representation of the *Plastaumatic* plastome assembly pipeline with details. There are six steps involved in the *Plastaumatic* pipeline i.e., trimming, creating config file for assembly, *de novo* assembly, standardization, annotation, and generating a feature table file. The pipeline can be run using Snakemake or a shell script, both of which produce the same output. The input data required for the *Plastaumatic* is the raw WGS reads in *fastq* format, a seed file in *fasta* format, and an annotation file in *GenBank* format. The pipeline creates multiple directories as detailed in the figure to store all the output.

(usually a well conserved gene), an annotated reference plastome in GenBank format, paths to the main executable script *NOVOPlasty*, path to the *Plastaumatic* repository, and a plastome assembly size range.

Execution

The pipeline can be executed by simply executing the *snakefile* from any desired directory. An additional wrapper script written in shell is also available for running the *Plastaumatic* pipeline and it does not require to have Snakemake installed. This can be executed by running *plastaumatic -s <seed.fa> -g <reference.gb> -r <range> -f <fof.txt> -n <NOVOPlasty4.3.1.pl>*. Here, paths to the raw sequencing data are provided in a simple text file (*fof.txt*). In both methods, upon execution a user-specified *prefix* directory is created for each sample and all the rules are run. Upon a successful run, links to plastome assembly and annotation of each sample are created under their *prefix* directory.

Pipeline

The first rule in *snakefile* is to perform adaptor removal and quality filtering using an ultra-fast FASTQ pre-processor *fastp* (Chen et al., 2018). We chose *fastp* because it provides faster performance and additional functionality such as automatic detection of adapter sequences and subsampling a fraction of reads from the input. We have tested different subsets for the filtered data (1, 5, 10, 15 million reads) and chose the optimal value of 10 million reads for this pipeline. All the other parameters are set to default for *fastp*. In the next rule, a config file required by *NOVOPlasty* is created by adding filtered read paths, path to seed file, range, and other parameters. A *de novo* plastome assembly is then performed using *NOVOPlasty* in the following rule. A successful *NOVOPlasty* run will create either a single circular assembly or two circular assemblies (Option 1 and 2) with different orientation of the SSC region. Also, the assemblies are not standardized, meaning the *fasta* file can start from any region in the plastome. Standardization of the assembly is necessary for accurate downstream analyses such as annotation and multiple plastome comparisons. Since no suitable publicly available tool was found for this process, a custom written shell script *standardize_cpDNA.sh* which uses *BLAST+* (Camacho et al., 2009) and *SAMtools* (Danecek et al., 2021) is used for standardization. The plastome assembly is first aligned with itself to get the repeat sequences and to locate the four main regions of a plastome (LSC, SSC, IRa, and IRb). The assembly is split at these four regions and joined together to make a standardized assembly in the form of LSC-IRb-SSC-IRa. If *NOVOPlasty* produces two assembly options, one option is selected based on the SSC orientation of the reference plastome and is used for subsequent analyses. In some rare cases, *NOVOPlasty* outputs ambiguous bases (non-ACTG) in the assembly. These are also corrected from the assembly using reads to get a final clean assembly. In the next rule, the standardized assembly is used for the annotation by *AnnoPlast.py*. Current plastome annotation tools either does not annotate some features or improperly annotate feature boundaries which require manual correction of these features. To overcome this issue, we have developed an annotation tool written in python called *AnnoPlast.py*. It uses *Blast+*, *Biopython* and *pandas* tools for annotation of target sequences from a reference GenBank file. First, all the features from the reference are extracted and then queried against the target sequence using *blastn*. Three rounds of *blastn* are carried out with different percent identities until all the features are mapped. Then the blast output is parsed and annotated to get the target annotations in GenBank format. The *AnnoPlast* annotation tool is also compared with the existing annotation tools such as *GeSeq* Tillich et al., 2017) and *PGA* (Qu et al., 2019). A common practice with newly assembled plastome sequences is to deposit them under the NCBI's GenBank database. A popular way of doing this is using BankIt which requires annotations in feature table format (suffix.*tbl*). In the final rule, a GenBank to tbl file conversion is carried out to get the feature table file. The progress of each rule is recorded and written into separate log files to debug any errors in the execution of the pipeline. This pipeline is also incorporated with the ability to assemble and annotate multiple plastomes by automatically creating all the required files for all sets of raw reads specified by the user. This would increase the efficiency when many plastomes are assembled.

Benchmarking

The automated pipeline was tested against three soybean genomes and twelve previously published potato genomes (Achakkagari et al., 2020; Kyriakidou et al., 2020). All soybean and potato plastomes were also manually assembled using the same programs used in the original study (Achakkagari et al., 2020). The protocol for manual assembly includes the command line execution of all the same programs. The timing for all manual or automated assemblies started before the first modification of any file or directory. The timing for all manual or automated assemblies stopped when the jobs finished. All outputs from three repetitions were checked for consistency before declaring the sets of assemblies as successful. All the steps were run on a machine with 180G memory and 16 threads. The parameters for the manual runs were kept same as the original study. The *Trimmomatic* parameters were set to *ILLUMINACLIP : TruSeq3-PE-2.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:60*. The parameters used for the *NOVOPlasty* are as follows: assembly

type - *chloro*, genome range - *120000 -200000* bp, k-mer size - *29*, max memory - *40G*, read length – *151 bp*, insert size – *300 bp*, and *PE* representing the type of reads as paired-end. The annotation performed by PGA were run with default parameters. The following versions of software were used in all the runs, *Trimmomatic v0.39*, *NOVOPlasty v4.3.1*, *SAMtools v1.13* and *BLAST+ v2.12.0, fastp v0.23.2*. The pipeline was also tested with more complex gymnosperms and angiosperms such as *Cryptocoryne elliptica* (Talkah et al., 2022), *Cyperus rotundus* (Wu et al., 2021b), and *Picea mariana* (Lo et al., 2020).

Results

Plastid genomes are an essential part of plant cells and play a fundamental role in photosynthesis. Plastomes are highly conserved, and characterization of their sequence help understand the evolutionary relationships among organisms. As more and more sequencing projects are on-going, our pipeline will help to speed up the analysis of plastome sequences. The *Plastaumatic* pipeline has integrated several publicly available and new tools to provide a complete analysis of plastome sequences. The raw reads are processed using *fastp* to solve the issue with finding and providing the adaptor sequences for trimming, since it automatically detects the adaptor sequences. Also, *fastp* has a functionality of subsampling raw reads to a specified amount, which greatly reduces the amount of data to be processed and improves speed. For the assembly, the most popular assembler, *NOVOPlasty*, is used. A new script was developed to solve common standardization issues with assembled plastome sequences. The scrip provides a standardized way of representing the plastome sequences, which is necessary for various downstream analyses. A new annotation tool to accurately annotate all the gene features in the target assembly was also developed. This is to overcome the issues with existing annotation tools improperly annotating gene boundaries or missing some features. And finally, the pipeline integrates a publicly available tool to get an annotation feature table file that is needed for submissions to NCBI's GenBank. In comparison to other similar pipelines for plastome analysis, *Plastaumatic*, provides more features and features that are essential in plastome analysis and characterization (Table 1).

The *Plastaumatic* pipeline can be executed either as a snakemake pipeline or as a shell script. Publicly available and novel sequences were used to test the *Plastaumatic* pipeline to determine its accuracy and efficiency. All twelve potato plastomes assembled by Plastaumatic were consistent with the manually assembled plastomes and their published plastome assemblies (Achakkagari et al., 2020). The three soybean plastomes assembled by the *Plastaumatic* were also consistent with the three soybean plastomes assembled manually (ON470217-ON470219). The *Cryptocoryne elliptica* assembly is consistent with its published assembly, whereas the *Cyperus rotundus* and *Picea mariana* assemblies have small differences compared with their previously published assemblies (Table 2). The *Picea mariana* assembly from the *Plastaumatic* has an additional insertion sequence of 28 bp, which likely resulted from different assembly methods used. Though the *Cyperus rotundus* assembly obtained from this study is longer than the published assembly, it is more consistent with plastomes from other *Cyperus* species. Hence it is highly likely that the original assembly of *Cyperus rotundus* is incomplete.

The annotations obtained from the *Plastaumatic* for each genome are the same as their original annotations. All the gene features were correctly annotated through *Plastaumatic*. The accuracy of annotations was also compared with the other available tools such as GeSeq and PGA. While the GeSeq performed better than PGA, it incorrectly annotated some gene features. The *rps12* gene is a trans-splicing gene and it is often difficult to annotate. GeSeq was unable to properly annotate the *rps12* gene and other genes such as *petB*, *petD*, and *rpl16*. The *PGA* tool also annotated the *rps12* gene incorrectly, along with *ycf3*, *ndhD* genes, and any *trnA* with a short length. Also, PGA does not report intron and exon features in the output GenBank file. In comparison to these tools, the *AnnoPlast* performs better and generates accurate gene features.

TABLE 1 Comparison of major features of different software for plastome analysis.

Feature	NOVOWrap	GetOrganelle	Fast-Plast	Plastaumatic
Trimming	X	X	✓	✓
de novo assembly	✓	✓	✓	✓
Standardization	✓	X	✓	✓
Annotation	X	X	X	✓
Feature table files	X	X	X	✓
Coverage plot	X	✓	✓	X

TABLE 2 List of species used in testing the pipeline.

Species	Taxonomy	SRA	GenBank	Size (original study)	Size (*Plastaumatic*)	Percent Identity
Solanum stenotomum subsp. *goniocalyx*	eudicots	SRR10244441	MT120855	155,492	155,492	100
Solanum stenotomum subsp. *goniocalyx*	eudicots	SRR10244440	MT120856	155,492	155,492	100
Solanum xajanhuiri	eudicots	SRR10244437	MT120857	155,486	155,486	100
Solanum phureja	eudicots	SRR10244439	MT120858	155,492	155,492	100
Solanum stenotomum subsp. *stenotomum*	eudicots	SRR10244438	MT120859	155,492	155,492	100
Solanum bukasovii	eudicots	SRR10244436	MT120860	155,491	155,491	100
Solanum juzepczukii	eudicots	SRR10248512	MT120863	155,532	155,532	100
Solanum chaucha	eudicots	SRR10248511	MT120864	155,518	155,518	100
Solanum tuberosum subsp. *andigena*	eudicots	SRR10248515	MT120861	155,530	155,530	100
Solanum tuberosum subsp. *andigena*	eudicots	SRR10248514	MT120862	155,518	155,518	100
Solanum tuberosum subsp. *tuberosum*	eudicots	SRR10248513	MT120865	155,564	155,564	100
Solanum curtilobum	eudicots	SRR10248510	MT120866	155,492	155,492	100
Cryptocoryne elliptica	monocots	SRR14784941	MZ435316.1	159,968	159,968	100
Picea mariana	conifers	SRR12885547	MT261462.1	123,961	123,986	99.97
Cyperus rotundus	monocots	SRR12799673	MT937176.1	182,986	186,127	100
*Glycine max**	eudicots	SRR19105742	ON470219	–	152,226	–
*Glycine max**	eudicots	SRR19103585	ON470217	–	152,226	–
*Glycine max**	eudicots	SRS6529047	ON470218	–	152,226	–

A table listing all the species used in this study to test the pipeline and their SRA and GenBank accession numbers. The plastome assembly size from the original study and this study are compared. Species marked with * are novel plastome assemblies generated using the Plastaumatic pipeline.

Twelve potato plastomes (for which plastome assemblies are previously published)

The time taken to assemble twelve potato plastomes manually was measured separately for each of the three repetitions to be 373, 281, and 336 minutes, respectively, resulting in an average time of 330 minutes with a peak memory usage of 40G. The time taken to assemble the twelve potato plastomes using the automated pipeline were 36, 36, and 36 minutes, resulting in an average assembly time of 36 minutes with a peak memory usage of 11G. The automated pipeline finished ~10x faster compared to the manual assembly with only ¼th of memory. Thus, *Plastaumatic* means a significant decrease in the time taken to finish the plastome assembly and annotations compared to the manual assembly.

Three soybean plastomes (not previously published)

The time taken to assemble three soybean plastomes manually, measured separately for each of the three repetitions, were 202, 215, and 182 minutes, respectively, resulting in an average time of 200 minutes with a peak memory usage of 40G. The time taken to assemble three soybean plastomes using the automated pipeline with three repetitions, were 28, 26, and 27 minutes resulting in an average assembly time of 27 minutes with a peak memory usage of 11G. Similar to the results with the potato genomes, the *Plastaumatic* finished ~7x faster compared to the manual assembly with about ¼th of memory. These soybean plastomes assemblies were submitted to the NCBI GenBank under the accession numbers ON470217-ON470219.

Discussion

Plastaumatic pipeline performance

For both the twelve potato genomes and the three soybean genomes we could confidently conclude that adopting the automated pipeline resulted in substantial decrease in time and memory needed for complete assembly and annotation, thus a huge increase in efficiency. The time taken for manual assembly were less consistent compared to the automated assembly using the pipeline. One of the factors contributing to such inconsistency was the amount manual inspection needed after steps such as annotation. In the twelve assembled and annotated potato plastomes, the *ycf3* gene feature was reported to contain internal stop codons. In all assembled and annotated soybean plastomes, the *ndhB* gene feature was reported to contain internal stop codons. Such results are due to error made by *PGA* during annotations. The detection of internal stop codons in assemblies usually calls for corrections made to the corresponding coordinates manually though *blastn* searches. As previously introduced, this adopted workflow of plastome

assembly consisted of six different programs to complete. Working on high-performance computing systems and executing each of the six programs would require various specifications, including but not limited to input and output full paths, containing directories, accessory references, or configurations. In the meantime, while the required user input information was overall not complicated, the forms of the paths or files requested by the six programs were not unified to optimize the compatibility of each tool to the others. Therefore, inputting information for each program would require more effort than copying and pasting the same texts from the previous step. When the number of genomes to be assembled is increased, the time needed for repeated inspections on whether the input information was correctly entered was substantially elevated. Such impact could be seen from the rather significant difference between the time taken for manual assembly of the three soybean plastomes and that of the twelve potato plastomes.

In all manual assemblies of plastomes, the user must wait for an upstream job, e.g., Trimmomatic, to finish before the downstream job, e.g., NOVOPlasty, could be submitted, since manual execution of each program requires indication of paths of the input files, which could not be known before the outputs were produced by the upstream program. Therefore, by the nature of doing manual plastome assemblies, some periods of time in between the execution of the programs would be wasted if the user did not get the notification message of an upstream program finishing, thus creating lag between connections and lengthening the overall time needed for manual plastome assemblies.

Limitations

Limitations to the pipeline are currently issues inherent from component software and are the same as with manual processing. For example, when the *de novo* assembly by *NOVOPlasty* is not finished properly, it yields multiple suggested assembled plastomes where the program was unable to decide which one is the most appropriate assembly. Since the number of outcomes from *NOVOPlasty* hardly exceeds two, the situation where more than two outcomes were produced by *NOVOPlasty* results the pipeline to exit. This can be controlled by providing a reference *fasta* sequence for *NOVOPlasty*, that is providing the path to a reference sequence in *fasta* format in the *NOVOPlasty* configuration file.

Comparison with existing tools

Overall, there has not been many attempts to automate the workflow of plastome assembly and annotation. One program with similar motivations was called *NOVOWrap* (Wu et al., 2021a). *NOVOWrap* was designed to automatically assemble, validate, and standardize plastomes with minimum inputs and user intervention. While such objectives sound similar to the *Plastaumatic* pipeline, they differ on many aspects. Firstly, with respect to the workflow of plastome assembly and annotation in the current study, the usage of *NOVOWrap* would only achieve partial automation. Users would still have to pre-process reads to remove adapter sequences to use as an input for *NOVOWrap*. Secondly, since *NOVOWrap* only provides functions of *de novo* assembly, validation and standardization, the final outputs of *NOVOWrap* would not be annotated, unlike the output of *Plastaumatic*, and would therefore require further processes before being NCBI publication ready. Finally, while both the *Plastaumatic* pipeline and *NOVOWrap* proposed automatic assembly of plastomes, our intentions were different enough to cause the two programs to be optimized in completely different directions. The ultimate aim for our automated pipeline was to eliminate manual input thus achieving higher efficiency in batch assembly and annotation of large numbers of plastomes altogether. Naturally, all parameters specified in our automated pipeline aimed to require the least amount of time when processing and is optimized to run on multiple genomes. In conclusion, while parts of our automated pipeline and *NOVOWrap* share similarities, they are optimized to perform very different types of assembly tasks, and *Plastaumatic* is well suited for complete plastome batch assembly and annotation with an NCBI-ready final output.

Data availability statement

Publicly available datasets were analyzed in this study. The sequencing data presented in the study are deposited in the NCBI repository (https://www.ncbi.nlm.nih.gov/bioproject/), the BioProject accession numbers are PRJNA556263, PRJNA835403, PRJNA627639, and PRJNA835489. The plastome assemblies presented in the study are deposited in the GenBank repository (https://www.ncbi.nlm.nih.gov/nuccore/), accession numbers are MT120855.1 - MT120866.1, and ON470217-ON470219.

Author contributions

WC, SA, and MS conceptualized the project, WC and SA developed the pipeline, drafted and edited the manuscript, MVS supervised the project and edited the manuscript. All authors contributed to the article and approved the submitted version.

Funding

Funding for the project was provided through a Compute/ Calcul Canada Resource Allocations for Research Portals and Platforms (The Potato Genome Diversity Portal) award and a Génome Québec award (GQ-AAC-2019-2) to MVS. Sequencing of the soybean plastomes was funded by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) to MVS.

Acknowledgments

The authors wish to thank Ilayda Bozan for technical help.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Achakkagari, S. R., Kyriakidou, M., Tai, H. H., Anglin, N. L., Ellis, D., and Strömvik, M.V. (2020). Complete plastome assemblies from a panel of 13 diverse potato taxa. *PloS One* 15 (10), e0240124. doi: 10.1371/journal.pone.0240124

Achakkagari, S. R., Tai, H. H., Davidson, C., Jong, H., and Strömvik, M. V. (2021). The complete plastome sequences of nine diploid potato clones. *Mitochondrial DNA B Resour.* 6 (3), 811–813. doi: 10.1080/23802359.2021.1883486

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30 (15), 2114–2120. doi: 10.1093/bioinformatics/btu170

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadoupoulos, J., Bealer, K., et al. (2009). BLAST+: Architecture and applications. *BMC Bioinf.* 10, 421. doi: 10.1186/1471-2105-10-421

Camargo Tavares, J. C., Achakkagari, S., Archambault, A., and Strömvik, M. V. (2022). The plastome of the arctic oxytropis arctobia (Fabaceae) has large differences compared with that of o. splendens and those of related species. *Genome* 65 (5), 301–313. doi: 10.1139/gen-2021-0059

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34 (17), i884–i890. doi: 10.1093/bioinformatics/bty560

Chung, H.-J., Jung, J. D., Park, H.-W., Kim, J.-H., Cha, H. W., Min, S. R., et al. (2006). The complete chloroplast genome sequences of solanum tuberosum and comparative analysis with solanaceae species identified the presence of a 241-bp deletion in cultivated potato chloroplast DNA sequence. *Plant Cell Rep.* 25 (12), 1369–1379. doi: 10.1007/s00299-006-0196-4

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., et al. (2021). *Twelve years of SAMtools and BCFtools* (Gigascience), Vol. 10, giab008. doi: 10.1093/gigascience/giab008

Dierckxsens, N., Mardulyn, P., and Smits, G. (2017). NOVOPlasty: *de novo* assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* 45 (4), e18. doi: 10.1093/nar/gkw955

Hoopes, G., Meng, X., Hamilton, J., Achakkagari, S. R., Alves Freitas Guesdes, F., Bolger, M. E., et al. (2022). Phased, chromosome-scale genome assemblies of tetraploid potato reveals a complex genome, transcriptome, and proteome landscape that underpin phenotypic diversity. *Mol. Plant* 15 (3), 520–536. doi: 10.1016/j.molp.2022.01.003

Jin, J. J., Yu, W. B., Yang, J. B., Song, Y., DePamphilis, C. W., Yi, T. S., et al. (2020). GetOrganelle: A fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. *Genome Biol.* 21, 241. doi: 10.1186/s13059-020-02154-5

Kyriakidou, M., Achakkagari, S. R., Galvez, J. H., Zhu, X., Tang, K., Tai, H., et al. (2020). Structural genome analysis in potato taxa. *Theor. Appl. Genet. (TAG)* 133, 951–966. doi: 10.1007/s00122-019-03519-6

Lo, T., Coombe, L., Lin, D., Warren, R. L., Kirk, H., Pandoh, P., et al. (2020). Complete chloroplast genome sequence of a black spruce (Picea mariana) from Eastern Canada. *Microbiol. Resour. Announcements* 9 (39), e00877–e00820. doi: 10.1128/MRA.00877-20

McCauley, D. E., Sundby, A. K., Bailey, M. F., and Welch, M. E. (2007). Inheritance of chloroplast DNA is not strictly maternal in silene vulgaris (Caryophyllaceae): evidence from experimental crosses and natural populations. *Am. J. Bot.* 94 (8), 1333–1337. doi: 10.3732/ajb.94.8.1333

McKain, M. R., and Wilson, M. (2017) *Fast-plast: rapid de novo assembly and finishing for whole chloroplast genomes*. Available at: https://github.com/mrmckain/Fast-Plast.

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., et al. (2021). Sustainable data analysis with snakemake [version 1; peer review: 1 approved, 1 approved with reservations]. *F1000Research* 10, 33. doi: 10.12688/f1000research.29032.1

Qu, X. J., Moore, M. J., Li, D. Z., and Yi, T. S. (2019). PGA: a software package for rapid, accurate, and flexible batch annotation of plastomes. *Plant Methods* 15, 50. doi: 10.1186/s13007-019-0435-7

Talkah, N. S. M., Wongso, S., and Othman, A. S. (2022). Complete chloroplast genome data for cryptocoryne elliptica (Araceae) from peninsular Malaysia. *Data Brief.* 42, 108075. doi: 10.1016/j.dib.2022.108075

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., et al. (2017). GeSeq – versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* 45, W6–W11. doi: 10.1093/nar/gkx391

Wu, P., Xu, C., Chen, H., Yang, J., Zhang, X., and Zhou, S. (2021a). NOVOWrap: An automated solution for plastid genome assembly and structure standardization. *Mol. Ecol. Resour.* 21 (6), 2177–2186. doi: 10.1111/1755-0998.13410

Wu, R., Yu, C., and Wu, Y. (2021b). Characterization of the complete plastome of cyperus rotundus l. (Cyperaceae). *Mitochondrial DNA Part B* 6 (1), 58–59. doi: 10.1080/23802359.2020.1845999

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 21 November 2022
DOI 10.3389/fpls.2022.1049209

OPEN ACCESS

EDITED BY
Weijun Kong,
Capital Medical University, China

REVIEWED BY
Long-Fei Fu,
Guangxi Institute of Botany (CAS), China
Zubaida Yousaf,
Lahore College for Women University, Pakistan
Yong Gao,
Qujing Normal University, China

*CORRESPONDENCE
Lan Wu
lwu@icmm.ac.cn
Li Xiang
lxiang@icmm.ac.cn

SPECIALTY SECTION
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

RECEIVED 20 September 2022
ACCEPTED 26 October 2022
PUBLISHED 21 November 2022

CITATION
Lan Z, Shi Y, Yin Q, Gao R, Liu C, Wang W, Tian X, Liu J, Nong Y, Xiang L and Wu L (2022) Comparative and phylogenetic analysis of complete chloroplast genomes from five *Artemisia* species. *Front. Plant Sci.* 13:1049209. doi: 10.3389/fpls.2022.1049209

Comparative and phylogenetic analysis of complete chloroplast genomes from five *Artemisia* species

Zhaohui Lan[1,2], Yuhua Shi[1], Qinggang Yin[1], Ranran Gao[1], Chunlian Liu[2], Wenting Wang[1], Xufang Tian[2], Jiawei Liu[3], Yiying Nong[3], Li Xiang[1]* and Lan Wu[1]*

[1]Key Laboratory of Beijing for Identification and Safety Evaluation of Chinese Medicine, Artemisinin Research Center, Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing, China, [2]College of Pharmacy, Hubei University of Chinese Medicine, Wuhan, China, [3]Department of product development, Hubei Aiaitie Health Technology Co., LTD, Huanggang, China

Artemisia Linn. is a large genus within the family Asteraceae that includes several important medicinal plants. Because of their similar morphology and chemical composition, traditional identification methods often fail to distinguish them. Therefore, developing an effective identification method for *Artemisia* species is an urgent requirement. In this study, we analyzed 15 chloroplast (cp) genomes, including 12 newly sequenced genomes, from 5 *Artemisia* species. The cp genomes from the five *Artemisia* species had a typical quadripartite structure and were highly conserved across species. They had varying lengths of 151,132–151,178 bp, and their gene content and codon preferences were similar. Mutation hotspot analysis identified four highly variable regions, which can potentially be used as molecular markers to identify *Artemisia* species. Phylogenetic analysis showed that the five *Artemisia* species investigated in this study were sister branches to each other, and individuals of each species formed a monophyletic clade. This study shows that the cp genome can provide distinguishing features to help identify closely related *Artemisia* species and has the potential to serve as a universal super barcode for plant identification.

KEYWORDS

***Artemisia* Linn., chloroplast genome, genome comparison, species identification, phylogenetic analysis, simple sequence repeat**

Abbreviations: IR, inverted repeat region; LSC, large single copy region; SSC, small single copy region; ML, maximum likelihood; rRNA, ribosomal RNA; SSR, simple sequence repeats; tRNA, transfer RNA; RSCU, relative synonymous codon usage.

Introduction

Artemisia Linn. is a large genus within the family Asteraceae comprising commonly used herbs that have a long history of medicinal use (Watson et al., 2002; Riggins, 2008; Kim G. B. et al., 2020). Among the various compounds present in these plants, terpenoids represent the main effective component. Modern pharmacological studies have shown that *Artemisia* medicinal plants exert diuretic, expectorant, antiinflammatory, hemostatic, hypotensive, and antiallergic effects (Chen et al., 2021; Hsueh et al., 2021; Hong et al., 2022). Owing to their similar morphological characteristics and chemical composition, *Artemisia* species are often mixed or substituted in different regions of China. According to the Chinese flora, *Artemisia princeps* and *A. lancea* are often mixed with *A. argyi*. It is difficult to distinguish dry herbs and raw materials using traditional identification methods. These difficulties have seriously hindered their development as medicinal plants. Some universal DNA barcodes, such as internal transcribed spacer (ITS) and ITS2, have been used to distinguish *Artemisia* species; however, these are inadequate for solving the classification problem because the sequences of closely related species are similar due to the hybridization of *Artemisia* plants (Garcia et al., 2008; Wang et al., 2016; Liu et al., 2017). Therefore, development of an accurate and effective method to identify medicinal *Artemisia* species is urgently needed.

The chloroplast (cp) is a multifunctional organelle with its independent genetic material. The structure of most angiosperm cp genomes is mostly conservative with a typical double-stranded, circular quadripartite structure, which includes a small single copy (SSC) region, large single copy (LSC) region, and two inverted repeat regions (IRa and IRb) (Jansen et al., 2005). With the development of sequencing technology, an increasing number of cp genomes have been published. The cp genome is usually 110–170 kb long, with 110–150 coding genes, which are highly conserved in gene type, gene number, and sequence compared with the mitochondrial or nuclear genome (Green, 2011; Zhu et al., 2016). The evolution rate of the cp genome is relatively moderate (Dong et al., 2013). Due to the lack of recombination, small genome size, and high single cell copy number, the cp genome is widely used in phylogenetic analysis and species identification (Dong et al., 2012; Twyford and Ness, 2017; Dong et al., 2017; Dong et al., 2018; Wu et al., 2018; Mader et al., 2018; Yang et al., 2018; Zhang et al., 2019). The comparison of cp genomes helps discover sequence variations, such as simple sequence repeats (SSRs), and mutation hotspots, and this has led some researchers to propose that the cp genome can be used as a super barcode for species identification (Li et al., 2015). Compared with the traditional relatively short and easily amplified DNA barcode, the cp genome has more abundant mutation site information and stronger species resolution ability, which can more accurately reflect the genetic characteristics of closely related species.

In this study, we used a second-generation sequencing platform to obtain the cp genomes from five *Artemisia* species. We compared their genome structure, codon usage preference, repeat sequences, and mutation hotspots. Finally, we performed a phylogenetic analysis of 29 cp genomes from 19 angiosperms. This study aimed to contribute valuable information toward the construction of the cp genome database of *Artemisia* species, which will aid in their identification.

Materials and methods

Sample collection, DNA extraction, and sequencing

Fifteen cp genomes from five *Artemisia* species were used in this study (Supplementary Table S1). Fresh leaves of 12 individuals from 5 *Artemisia* species were collected from Hainan, Hubei, and Beijing in China. The cp genomes of two additional individuals were downloaded from NCBI (Accession No.: MZ151340.1 and MW411453.1, *A. lactiflora*) and one cp genome was obtained from our previously published study (Accession No.: ON381734, *A. indica*) (Lan et al., 2022). The genomic DNA of each individual was extracted from fresh leaves using the plant DNA Extraction Kit (QIAGEN, Germany). The quality and concentration of genomic DNAs were evaluated using the Qubit2.0 Fluorometer (Thermo Scientific, USA) and NanoDrop 2000c spectrophotometer (Nanodrop Technologies, Wilmington, DE, USA) to ensure they met the requirements for sequencing. Based on the Illumina Nova Seq sequencing platform, 2 × 150 bp sequencing was performed with a depth of 226–578×. After quality pruning, clean reads were obtained from the original sequencing data for subsequent splicing and annotation.

Genome assembly and annotation

The NOVOPlasty software (https://github.com/ndierckx/NOVOPlasty) was used to assemble the cp genomes. We compared the clean reads with the scaffold obtained from the assembly, optimized the assembly results according to the paired-end and overlap relationships of the reads, and used the GapCloser software (v1.12, http://soap.genomics.org.cn/soapdenovo.html) to repair the inner hole of the assembly result. Finally, the reference genome was used to correct the starting position of the assembled cp sequence and determine the position and direction of four cp partitions (LSC/IRa/SSC/IRb) for obtaining the final cp genome sequence. The cp genomes were annotated using CpGAVAS (Liu et al., 2012). The genome circle

maps were drawn using the online tool OGDRAWH (http://ogdraw.mpimp-golm.mpg.de/) (Lohse et al., 2007). The cp genomes and gene annotation files were uploaded to the NCBI database to obtain GenBank accession numbers.

Codon usage analysis

The relative synonymous codon usage (RSCU) of the 15 cp genomes from the 5 *Artemisia* species was determined and analyzed using the CodonW1.4.2 software (http://mobyle.pasteur.fr/cgi-bin/portal.py?form=codonw). Heat maps were constructed using the RSCU values. An RSCU value of >1 indicates that the codon is used more frequently, a value equal to 1 indicates that the codon has no usage preference, and a value of <1 indicates that the codon is used less frequently.

Repeat sequences and simple sequence repeat analysis

Four types of repeat sequences—forward, reverse, complementary, and palindromic—were identified using REPuter with a Hamming distance of 3 and a minimum repeat size of 30 bp (Kurtz et al., 2001). SSRs were detected using MISA with the following parameters: eight repeat units for mononucleotides; four for di- and trinucleotides; and three for tetra-, penta-, and hexanucleotides (Thiel et al., 2003; Lin et al., 2012).

Nucleotide diversity analysis

The nucleotide diversity of the 15 cp genomes was calculated *via* sliding window analysis using the DnaSP v5.10 software. The window length was set to 600 bp and the step length to 200 bp (Zhang et al., 2021).

Phylogenetic analysis

The phylogenetic tree was constructed based on the 15 whole cp genomes from 5 *Artemisia* species, another 13 species, and 1 outgroup *Cirsium japonicum*. All genomes, expect the 12 newly sequenced genomes, were downloaded from NCBI. The 29 cp genomes were compared using the MAFFT software (http://mafft.cbrc.jp/alignment/software/). Phylogenetic analysis was performed using the maximum likelihood (ML) method. Using IQtree's default parameters, ModelFinder automatically filters the best model to build the ML tree (Bootstrap to 1000) (Nguyen et al., 2015).

Results

Chloroplast genome sequencing and features of five *Artemisia* species

The cp genome lengths of *A. lancea*, *A. princeps*, *A. lactiflora*, *A. indica*, and *A. argyi* were 151,132 bp, 151,154 bp, 151,178 bp, 151,161 bp, and 151,152 bp, respectively (Table 1 and Supplementary Table S1). These genomes have a typical quadripartite structure, including an LSC (82,870–82,911 bp), SSC (18,338–18,354 bp), and two IR (24,960–24,961 bp) regions (Table 1). The average GC content was 37.5%, and the IR regions possessed higher GC content (43.1%) than the LSC (35.5%–35.6%) and SSC (30.9%) regions. In this study, we annotated 132 genes from the 15 cp genomes, of which 7 tRNA genes, 4 rRNA genes, and 7 protein-coding genes were duplicated in the IR

TABLE 1 Basic cp genome information of five *Artemisia* species.

Characteristics	*A. lancea*	*A. princeps*	*A. lactiflora*	*A. indica*	*A. argyi*
Raw data no.	39,907,620	39,570,956	35,584,122	31,794,994	34,216,492
Chloroplast genome coverage (×)	246	226	325	256	578
Total size (bp)	151,132	151,154	151,178	151,161	151,152
LSC length (bp)	82,870	82,880	82,911	82,901	82,891
IR length (bp)	24,960	24,960	24,960	24,961	24,960
SSC length (bp)	18,342	18,354	18,347	18,338	18,341
Total genes	132	132	132	132	132
Protein coding genes	87	87	87	87	87
tRNA genes	37	37	37	37	37
rRNA genes	8	8	8	8	8
Overall GC content (%)	37.5	37.5	37.5	37.5	37.5
GC content in LSC (%)	35.6	35.5	35.5	35.5	35.6
GC content in IR (%)	43.1	43.1	43.1	43.1	43.1
GC content in SSC (%)	30.9	30.9	30.9	30.9	30.9

region. A total of 114 genes were unique, including 80 protein-coding, 30 tRNA, and 4 rRNA genes.

Relative synonymous codons usage

The cp genomes of the 5 *Artemisia* species contained 64 codons. Of these, 61 codons encoded 20 proteins, and the other 3 were termination codons. The codon AUU had the highest usage frequency (1,078–1,079) with an RSCU value of 1.46–1.47. The RSCU values of the cp genomes in the five species were slightly different. Methionine (Met) and tryptophan (Trp) were encoded by a single codon, with an RSCU value of 1, indicating no preference. All other amino acids were encoded by multiple codons (Figure 1 and Supplementary Table S2). Arg, Leu, and Ser were encoded by six codons; alanine (Ala), glycine (Gly), proline (Pro), threonine (Thr), and valine (Val) by four codons; isoleucine (Ile) by three codons; and the rest were encoded by two codons.

Repeat and simple sequence repeat analyses

Some repeats with a length of ≥30 bp are known as long repeats. These are conducive to cp genome rearrangements and increase the genetic diversity of the population. In total, we found 42–50 long repeats in the 15 cp genomes of the 5 *Artemisia* species, including 19–22 forward, 20–22 palindromic, and 3–6 reverse repeats. Most of these repeats, which were 30–39 bp long, were located in the gene spacer and intron regions. This length of repeats was dominant in the *Artemisia* cp genomes, and the longest was the forward repeat. Complementary repeats were not identified in these genomes (Figure 2).

SSRs mainly comprise 1–6 types of nucleotide repeats. The cp genome exhibits characteristics of parthenogenesis, and SSRs highly vary within the same species. Therefore, SSR is widely used as a molecular marker in genetic map construction, target gene calibration, and mapping. We observed a total of 189–192 SSRs in the cp genomes of the 5 *Artemisia* species. Of these, 118–121 were mononucleotide SSRs, and most of them were of the A/T type (Figure 3 and Supplementary Table S3). The numbers of di-, tri-, tetra-, penta-, and hexanucleotide SSRs were 50–51, 5–6, 14–15, 1–2, and 0–1, respectively. Using comparative analysis, we found that the five *Artemisia* species had similar SSRs; however, the pentanucleotide repeat AAAAT/ATTTT only existed in *A. argyi* and the hexanucleotide repeat AAATAT/ATATTT only existed in *A. indica*. Overall, most SSRs comprised mono- or dinucleotide repeats. The types of oligonucleotide repeats were rich (Figure 3).

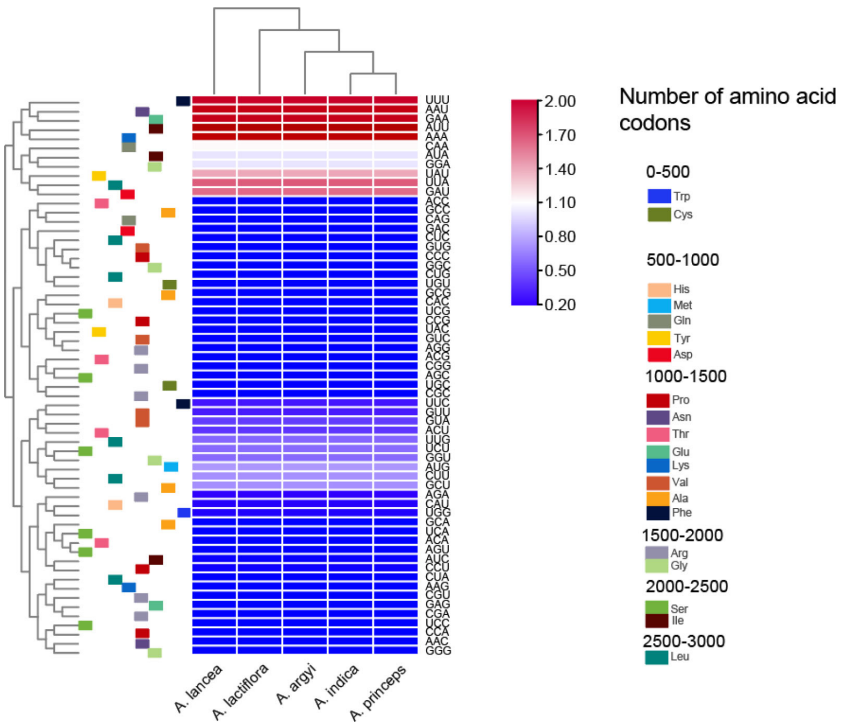


FIGURE 1
Heat map of the relative synonymous codons usage values of the cp genomes in the five *Artemisia* species.

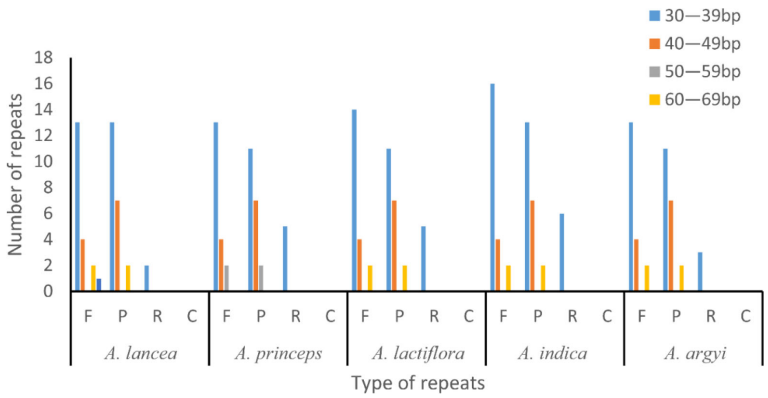


FIGURE 2
Long repeat sequence analysis of the genomes of five *Artemisia* species. F, forward repeat; P, palindromic repeat; R, reverse repeat; C, complementary repeat.

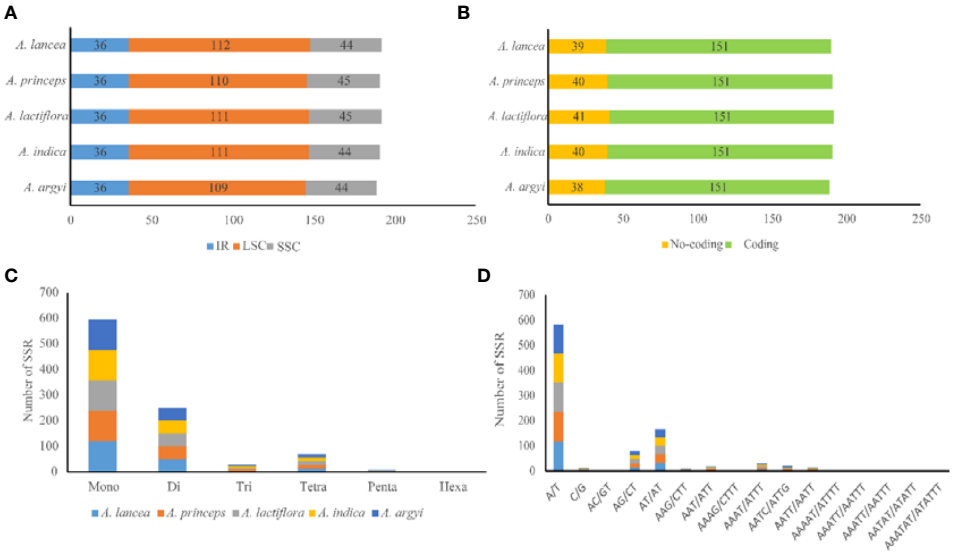


FIGURE 3
Type and distribution of SSRs in the five *Artemisia* cp genomes. **(A)** Frequency of SSRs in the LSC, SSC, and IR regions. **(B)** SSR distribution between coding and noncoding regions. **(C)** Number of SSR types. **(D)** Number of identified SSR motifs in different repeat class types. SSR, simple sequence repeat; LSC, large single copy region; SSC, small single copy region; IR, inverted repeat region.

Comparative analysis of the Cp genome

We used the DnaSP software to compare the nucleotide variation values (Pi) between all genes and intergenic regions in the cp genomes of the five *Artemisia* species. The hypervariable regions were detected, and the sequence differences were analyzed. Sliding window analysis revealed that the nucleotide diversity values within 600 bp varied from 0 to 0.006. Four mutational hotspots in the LSC and SSC regions were identified, including *rpl32_trnL-UAG*, *trnY-GUA_trnE-UUC*, *ndhH_rps15*, and *ycf1* (Figure 4). These can be used as potential sites for studying population genetics and the identification of *Artemisia* species.

Phylogenetic analysis

We constructed an ML tree using 29 cp genomes: 15 from the 5 *Artemisia* species used in this study and others from

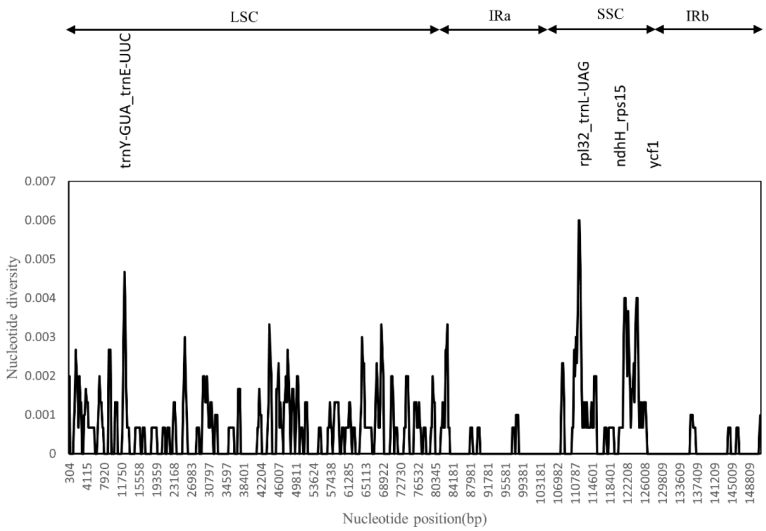


FIGURE 4
Sliding window test of nucleotide diversity (Pi) in the multiple alignments of the five *Artemisia* cp genomes. Peak regions with a Pi value of >0.004 were labeled with loci tags of the genic or intergenic region names. Pi values were calculated in the 600 bp sliding windows with steps of 200 bp. LSC, large single copy region; IRa, inverted repeat region a; SSC, small single copy region; IRb, inverted repeat region b.

another 13 species and 1 outgroup. We found that all *Artemisia* species clustered together, and different repeat individuals in each species formed a monophyletic branch with a high branch supporting rate, indicating that the cp genome could distinguish the five *Artemisia* species. *A. lactiflora* showed the closest relationship to *A. princeps*, followed by *A. indica* and *A. argyi*, and was distant from *A. lancea* (Figure 5). *A. scoparia* and *A. ordosica* were grouped into one branch, revealing a close relationship between them.

Discussion

In this study, we reported 12 newly sequenced cp genomes from 5 *Artemisia* medicinal species. We found that the genomes were extremely similar, with their size ranging from 151,132–151,178 bp. They belonged to medium-sized cp genomes compared to other Asteraceae species. The cp genomes of five *Artemisia* species contain 114 genes, which was similar to those of *Artemisia annua* (Shen et al., 2017). Like most other *Artemisia*

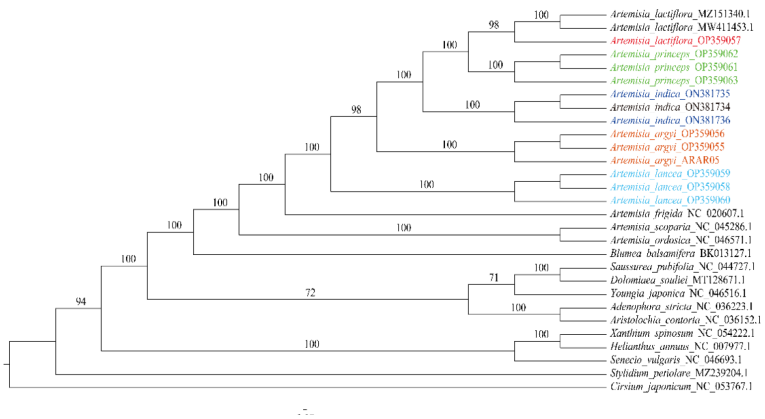


FIGURE 5
Phylogenetic tree constructed using the ML method based on the 29 cp genomes from 19 species. The numbers above the branches represent the ML bootstrap values. ML, maximum likelihood.

species, ycf1 and rps19 were also detected, but the copy number and location were different, ycf1 spans the IRb/SSC boundary, this was also seen in *A. scoparia* and *A. absinthium* (Chen et al., 2022). The GC content of the IR region was significantly greater than that of the LSC and SSC regions. The AT content was higher than the GC content in all cp genomes. As observed in most plants, we found that the cp genome of *Artemisia* was conservative and no rearrangements were detected in the five species. Multiple codons that encode the same amino acid are known as synonymous codons. Codon usage is unequal, as some synonymous codons are used more frequently than others, a phenomenon known as codon preference. Codon preferences develop in the long-term evolution of plants, and different species have distinct preferences. In this study, we found that the amino acid Leu had the highest proportion of codons in the cp genomes of the five *Artemisia* species.

SSRs widely exist in the cp genome and provide important information regarding population genetics and evolution. Their types, numbers, and distribution vary in each plant. In this study, 57.29%–58.33% of the SSRs were mapped to the LSC region. An SSR-rich region may harbor mutational hotspots (George et al., 2015). The A/T type accounted for the largest proportion of SSRs. The obvious nucleotide bias may be due to the lower number of hydrogen bonds and lower energy consumption of A/T bases (Niu et al., 2017; Kim and Cheon, 2021). Previous studies have reported a higher A/T than G/C content in most plants, which may be due to the large number of A/T-type SSRs (Wang et al., 2021; Wu et al., 2021; Han et al., 2022). In this study, we analyzed the number, location, and composition of SSRs in the cp genomes of five *Artemisia* species and provided a new reference for further research on molecular markers, mutation hotspots, population genetics, and crop breeding.

The mutation hotspot analysis revealed a high degree of similarity among the cp genomes of the five *Artemisia* species, implying that the differentiation of these species was lower than that of other species. The lack of genome information has hindered the classification, identification, and protection of *Artemisia* species. The cp genome sequence provides a basis for the further study on genome evolution and the development of genetic resources. Mutation hotspots are often used for species identification, and these highly variable regions can serve as specific DNA barcodes. In this study, we identified four hypervariable regions—*rpl32_trnL-UAG*, *trnY-GUA_trnE-UUC*, *ndhH_rps15*, and *ycf1*—all of which have the potential to be used as DNA barcodes for subsequent studies on *Artemisia* species.

Phylogenetic analysis is extremely important for clarifying the genetic relationship between species and for protecting, rationally developing, and utilizing plant resources. The cp genome can solve some issues that morphological taxonomy cannot; hence, it has widely been used to explore the phylogenetic relationships between species (Kim Y. K. et al., 2020; Zhao et al., 2021). Due to the low degree of genetic differentiation and similar morphology of *Artemisia* species, obtaining more information on the genetic features of *Artemisia* is expected to improve phylogenetic resolution. In this study, using phylogenetic analysis, we showed that *Artemisia* was a branch of Asteraceae, the five *Artemisia* species are sister groups that can distinguish each other and different repeat individuals in each *Artemisia* species formed a monophyletic branch, indicating that the cp genome can be used as a super barcode to distinguish the five *Artemisia* species. The cp genome provided an effective marker for inferring the phylogenetic relationships between *Artemisia* species.

Conclusion

In this study, we analyzed 15 cp genomes from 5 *Artemisia* species, including 12 newly sequenced genomes from *A. argyi*, *A. lactiflora*, *A. indica*, *A. princeps*, and *A. lancea*, all of which have been used as medicinal plants for a long time. The cp genomes were similar in structure and gene content and were highly conserved. Four hotspot regions and 189–192 SSR molecular markers were identified, which can serve as potential DNA barcodes for further studies on *Artemisia* species. The phylogenetic analysis showed that the entire cp genome provides distinguishing features to help identify the five *Artemisia* species with high support rates. This study will contribute to the study of population genetics, species identification, and conservation biology of *Artemisia* species.

Data availability statement

The data presented in the study are deposited in the NCBI repository, accession numbers OP359055–63, and ON381735–36.

Author contributions

LW and LX designed the research study. ZL and RG performed the research. LW, QY, YS, JL, and YN collected *Artemisia* plant materials. ZL, CL, WW, and XT analyzed the data. ZL and LW wrote and revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the National Natural Science Foundation of China (U1812403-1 and 81903758), the CACMS Innovation Fund (CI2021A05103 and CI2021A04112), the Fundamental Research Funds for the Central public welfare research institutes (ZZ13-YQ-106).

Acknowledgments

I would like to thank HT, GD, and TW for their valuable guidance during this research.

Conflict of interest

Authors JL and YN were employed by the company Hubei Aiaitie Health Technology Co., LTD.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1049209/full#supplementary-material

SUPPLEMENTARY TABLE 1
The cp genome information of the five *Artemisia* species.

SUPPLEMENTARY TABLE 2
Relative synonymous codon usage values of the five *Artemisia* species.

SUPPLEMENTARY TABLE 3
Simple sequence repeat types in the five *Artemisia* species.

References

Chen, C., Miao, Y., Luo, D., Li, J., Wang, Z., Luo, M., et al. (2022). Sequence characteristics and phylogenetic analysis of the *Artemisia argyi* chloroplast genome. *Front. Plant Sci.* 13, 906725. doi: 10.3389/fpls.2022.906725

Chen, J., Xu, X., Lin, L., Guo, D., Gui, W., Lin, Y., et al. (2021). Research progress on pharmacological effects of artemisia argyi. *J. Pharm. Res.* 40, 5. doi: 10.13506/j.cnki.jpr.2021.12.009

Dong, W., Liu, J., Yu, J., Wang, L., and Zhou, S. (2012). Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. *PloS One* 7, e35071. doi: 10.1371/journal.pone.0035071

Dong, W., Xu, C., Cheng, T., Lin, K., and Zhou, S. (2013). Sequencing angiosperm plastid genomes made easy: A complete set of universal primers and a case study on the phylogeny of saxifragales. *Genome Biol. Evol.* 5, 989–997. doi: 10.1093/gbe/evt063

Dong, W., Xu, C., Li, W., Xie, X., Lu, Y., Liu, Y., et al. (2017). Phylogenetic resolution in juglans based on complete chloroplast genomes and nuclear DNA sequences. *Front. Plant Sci.* 8, 1148. doi: 10.3389/fpls.2017.01148

Dong, W., Xu, C., Wu, P., Cheng, T., Yu, J., Zhou, S., et al. (2018). Resolving the systematic positions of enigmatic taxa: Manipulating the chloroplast genome data of saxifragales. *Mol. Phylogenet. Evol.* 126, 321–330. doi: 10.1016/j.ympev.2018.04.033

Garcia, S., Canela, M. A., Garnatje, T., Mcarthur, E. D., Pellicer, J., Sanderson, S. C., et al. (2008). Evolutionary and ecological implications of genome size in the north American endemic sagebrushes and allies (Artemisia, asteraceae). *Biol. J. Linn. Soc.* 94, 631–649. doi: 10.1111/j.1095-8312.2008.01001.x

George, B., Bhatt, B. S., Awasthi, M., George, B., and Singh, A. K. (2015). Comparative analysis of microsatellites in chloroplast genomes of lower and higher plants. *Curr. Genet.* 61, 665–677. doi: 10.1007/s00294-015-0495-9

Green, B. R. (2011). Chloroplast genomes of photosynthetic eukaryotes. *Plant J.* 66, 34–44. doi: 10.1111/j.1365-313X.2011.04541.x

Han, H., Qiu, R., Liu, Y., Zhou, X., Gao, C., Pang, Y., et al. (2022). Analysis of chloroplast genomes provides insights into the evolution of agropyron. *Front. Genet.* 13, 832809. doi: 10.3389/fgene.2022.832809

Hong, F., Zhao, M., Xue, L. L., Ma, X., Liu, L., Cai, X. Y., et al. (2022). The ethanolic extract of artemisia anomala exerts anti-inflammatory effects *via* inhibition of NLRP3 inflammasome. *Phytomedicine* 102, 154163. doi: 10.1016/j.phymed.2022.154163

Hsueh, T. P., Lin, W. L., Dalley, J. W., and Tsai, T. H. (2021). The pharmacological effects and pharmacokinetics of active compounds of artemisia capillaris. *Biomedicines* 10, 1412. doi: 10.3390/biomedicines9101412

Jansen, R. K., Raubeson, L. A., Boore, J. L., Depamphilis, C. W., Chumley, T. W., Haberle, R. C., et al. (2005). Methods for obtaining and analyzing whole chloroplast genome sequences. *Methods Enzymol.* 395, 348–384. doi: 10.1016/S0076-6879(05)95020-9

Kim, K. A., and Cheon, K. S. (2021). Complete chloroplast genome sequence of adenophora racemosa *(Campanulaceae)*: Comparative analysis with congeneric species. *PloS One* 16, e0248788. doi: 10.1371/journal.pone.0248788

Kim, Y. K., Jo, S., Cheon, S. H., Joo, M. J., Kim, K. J., Hong, J. R., et al. (2020). Plastome evolution and phylogeny of orchidaceae, with 24 new sequences. *Front. Plant Sci.* 11, 22. doi: 10.3389/fpls.2020.00022

Kim, G. B., Lim, C. E., Kim, J. S., Kim, K., Lee, J. H., Yu, H. J., et al. (2020). Comparative chloroplast genome analysis of artemisia *(Asteraceae)* in East Asia: Insights into evolutionary divergence and phylogenomic implications. *BMC Genomics* 21, 415. doi: 10.1186/s12864-020-06812-7

Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. (2001). REPuter: The manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* 29, 4633–4642. doi: 10.1093/nar/29.22.4633

Lan, Z., Tian, X., Shi, Y., Gao, R., Yin, Q., Xiang, L., et al. (2022). Chloroplast genome structure characteristics and phylogenetic analysis of artemisia indica. *China J. Chin. Materia. Med.*, 47, 224–231. doi: 10.19540/j.cnki.cjcmm.20220713.101

Lin, C. P., Wu, C. S., Huang, Y. Y., and Chaw, S. M. (2012). The complete chloroplast genome of ginkgo biloba reveals the mechanism of inverted repeat contraction. *Genome Biol. Evol.* 4, 374–381. doi: 10.1093/gbe/evs021

Liu, G., Ning, H., Ayidaerhan, N., and Aisa, H. A. (2017). Evaluation of DNA barcode candidates for the discrimination of artemisia l. *Mitochondrial. DNA Part A.* 28, 956–964. doi: 10.1080/24701394.2016.1219729

Liu, C., Shi, L., Zhu, Y., Chen, H., Zhang, J., Lin, X., et al. (2012). CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. *BMC Genomics* 13, 715. doi: 10.1186/1471-2164-13-715

Li, X., Yang, Y., Henry, R. J., Rossetto, M., Wang, Y., and Chen, S. (2015). Plant DNA barcoding: from gene to genome. *Biol. Rev. Camb. Philos. Soc* 90, 157–166. doi: 10.1111/brv.12104

Lohse, M., Drechsel, O., and Bock, R. (2007). OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Curr. Genet.* 52, 267–274. doi: 10.1007/s00294-007-0161-y

Mader, M., Pakull, B., Blanc-Jolivet, C., Paulini-Drewes, M., Bouda, Z. H., Degen, B., et al. (2018). Complete chloroplast genome sequences of four meliaceae species and comparative analyses. *Int. J. Mol. Sci.* 19, 701. doi: 10.3390/ijms19030701

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Niu, Z., Xue, Q., Wang, H., Xie, X., Zhu, S., Liu, W., et al. (2017). Mutational biases and GC-biased gene conversion affect GC content in the plastomes of dendrobium genus. *Int. J. Mol. Sci.* 18, 2307. doi: 10.3390/ijms18112307

Riggins, C. (2008). *Molecular phylogenetic and biogeographic study of the genus artemisia (Asteraceae), with an emphasis on section absinthium [D]* (USA: University of Illinois at Urbana-Champaign).

Shen, X., Wu, M., Liao, B., Liu, Z., Bai, R., Xiao, S., et al. (2017). Complete chloroplast fenome sequence and phylogenetic analysis of the medicinal plant artemisia annua. *Molecules* 22, 1330. doi: 10.3390/molecules22081330

Thiel, T., Michalek, W., Varshney, R. K., and Graner, A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare l.). *Theor. Appl. Genet.* 106, 411–422. doi: 10.1007/s00122-002-1031-0

Twyford, A. D., and Ness, R. W. (2017). Strategies for complete plastid genome sequencing. *Mol. Ecol. Resour.* 17, 858–868. doi: 10.1111/1755-0998.12626

Wang, Y., Wang, S., Liu, Y., Yuan, Q., Sun, J., and Guo, L. (2021). Chloroplast genome variation and phylogenetic relationships of atractylodes species. *BMC Genomics* 22, 103. doi: 10.1186/s12864-021-07394-8

Wang, X., Zheng, S., Liu, Y., and Han, J. (2016). ITS2, a better DNA barcode than ITS in identification of species in artemisia l. *Chin. Herbal. Medicines* 8, 352–358. doi: 10.1016/S1674-6384(16)60062-X

Watson, L. E., Bates, P. L., Evans, T. M., Unwin, M. M., and Estes, J. R. (2002). Molecular phylogeny of subtribe artemisiinae (Asteraceae), including artemisia and its allied and segregate genera. *BMC Evol. Biol.* 2, 17. doi: 10.1186/1471-2148-2-17

Wu, M. L., Li, Q., Xu, J., and Li, X. W. (2018). Complete chloroplast genome of the medicinal plant amomum compactum: gene organization, comparative analysis, and phylogenetic relationships within zingiberales. *Chin. Med.* 13, 10. doi: 10.1186/s13020-018-0164-2

Wu, L., Wu, M., Cui, N., Xiang, L., Li, Y., Li, X., et al. (2021). Plant super-barcode: a case study on genome-based identification for closely related species of fritillaria. *Chin. Med.* 16, 52. doi: 10.1186/s13020-021-00460-z

Yang, Z., Zhao, T., Ma, Q., Liang, L., and Wang, G. (2018). Comparative genomics and phylogenetic analysis revealed the chloroplast genome variation and interspecific relationships of corylus (Betulaceae) species. *Front. Plant Sci.* 9, 927. doi: 10.3389/fpls.2018.00927

Zhang, L., Wang, S., Su, C., Harris, A. J., Zhao, L., Su, N., et al. (2021). Comparative chloroplast genomics and phylogenetic analysis of zygophyllum (Zygophyllaceae) of China. *Front. Plant Sci.* 12, 723622. doi: 10.3389/fpls.2021.723622

Zhang, T., Xing, Y., Xu, L., Bao, G., Zhan, Z., Yang, Y., et al. (2019). Comparative analysis of the complete chloroplast genome sequences of six species of pulsatilla miller, ranunculaceae. *Chin. Med.* 14, 53. doi: 10.1186/s13020-019-0274-5

Zhao, F., Chen, Y. P., Salmaki, Y., Drew, B. T., Wilson, T. C., Scheen, A. C., et al. (2021). An updated tribal classification of lamiaceae based on plastome phylogenomics. *BMC Biol.* 19, 2. doi: 10.1186/s12915-020-00931-z

Zhu, A., Guo, W., Gupta, S., Fan, W., and Mower, J. P. (2016). Evolutionary dynamics of the plastid inverted repeat: The effects of expansion, contraction, and loss on substitution rates. *New Phytol.* 209, 1747–1756. doi: 10.1111/nph.13743

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 09 December 2022
DOI 10.3389/fpls.2022.1043740

OPEN ACCESS

EDITED BY
Tapan Kumar Mohanta,
University of Nizwa, Oman

REVIEWED BY
Gopal Pandi,
Madurai Kamaraj University, India
Krishnaveni Muthan,
Manonmaniam Sundaranar University, India
Ravendran Vasudevan,
University of Cambridge, United Kingdom

*CORRESPONDENCE
SeonJoo Park
sjpark01@ynu.ac.kr
Gi-Heum Nam
namgih@korea.kr

SPECIALTY SECTION
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

RECEIVED 14 September 2022
ACCEPTED 07 November 2022
PUBLISHED 09 December 2022

CITATION
Raman G, Nam G-H and Park S (2022) Extensive reorganization of the chloroplast genome of *Corydalis platycarpa*: A comparative analysis of their organization and evolution with other *Corydalis* plastomes. *Front. Plant Sci.* 13:1043740. doi: 10.3389/fpls.2022.1043740

Extensive reorganization of the chloroplast genome of *Corydalis platycarpa*: A comparative analysis of their organization and evolution with other *Corydalis* plastomes

Gurusamy Raman[1], Gi-Heum Nam[2*] and SeonJoo Park[1*]

[1]Department of Life Sciences, Yeungnam University, Gyeongsan, Gyeongsan-buk, Republic of Korea, [2]Plants Resource Division, Biological Resources Research Department, National Institute of Biological Resources, Seo-gu, Incheon, Republic of Korea

Introduction: The chloroplast (cp) is an autonomous plant organelle with an individual genome that encodes essential cellular functions. The genome architecture and gene content of the cp is highly conserved in angiosperms. The plastome of *Corydalis* belongs to the Papaveraceae family, and the genome is comprised of unusual rearrangements and gene content. Thus far, no extensive comparative studies have been carried out to understand the evolution of *Corydalis* chloroplast genomes.

Methods: Therefore, the *Corydalis platycarpa* cp genome was sequenced, and wide-scale comparative studies were conducted using publicly available twenty *Corydalis* plastomes.

Results: Comparative analyses showed that an extensive genome rearrangement and IR expansion occurred, and these events evolved independently in the *Corydalis* species. By contrast, the plastomes of its closely related subfamily Papaveroideae and other Ranunculales taxa are highly conserved. On the other hand, the synapomorphy characteristics of both *accD* and the *ndh* gene loss events happened in the common ancestor of the *Corydalis* and sub-clade of the *Corydalis* lineage, respectively. The *Corydalis*-sub clade species (*ndh* lost) are distributed predominantly in the Qinghai-Tibetan plateau (QTP) region. The phylogenetic analysis and divergence time estimation were also employed for the *Corydalis* species.

Discussion: The divergence time of the *ndh* gene in the *Corydalis* sub-clade species (44.31 – 15.71 mya) coincides very well with the uplift of the Qinghai-Tibet Plateau in Oligocene and Miocene periods, and maybe during this period, it has probably triggered the radiation of the *Corydalis* species.

Conclusion: To the best of the authors' knowledge, this is the first large-scale comparative study of *Corydalis* plastomes and their evolution. The present study may provide insights into the plastome architecture and the molecular evolution of *Corydalis* species.

KEYWORDS

Corydalis, Plastome rearrangement, relocation, IR expansion, *accD*, *clpP*, *ndh*, divergent time

Introduction

In angiosperms, chloroplast (cp) genomes are highly conserved in terms of their structure, gene content, and gene arrangement contains a pair of inverted repeats (IRs) that separate with a large single-copy (LSC) and a small single-copy (SSC) region (Palmer, 1983; Palmer, 1985; Wicke et al., 2011; Maliga, 2014; Lee et al., 2016; Mower and Vickrey, 2018). The cp genome of angiosperms comprises roughly 80 protein-coding genes, which play a role in essential cellular functions and photosynthesis, along with 30 transfer and four ribosomal RNA genes (Bock, 2007). Among these, approximately seventeen genes were duplicated in the IR region. In addition, most of the land plant cp genome size varies from 110 to 170 kb, and the difference in cp size is frequently ascribed to extension, reduction, or loss of the IR region (Chumley et al., 2006; Wicke et al., 2011). The large-scale IR expansion is identified in the *Pelargonium transvallense* (Geraniaceae) (Weng et al., 2014), in which the IR enlarged higher than tripled (87.7 kb) compared to the typical size of the IR region (~25 kb). On the other hand, the IR region of two lineages of *Erodium* (Geraniaceae) (Blazier et al., 2016; Ruhlman et al., 2017), *Carnegiea gigantean* (Cactaceae) (Sanderson et al., 2015), *Tahina spectabilis* (Arecaceae) (Choi et al., 2019), the Putranjivoid clade of Malpighiales (Jin et al., 2020a), and IR-lacking clade (IRLC) of Papilionoideae (Fabaceae) (Palmer and Thompson, 1982) plastome size reduced significantly. Generally, gene arrangement is not often in most angiosperms plastomes (Frailey et al., 2018). If so, the plastome rearrangement is relatively small (Xu and Wang, 2021). Nevertheless, a large amount of rearrangement is rare, but it is infrequently present in a few lineages, namely Asteraceae (Jansen and Palmer, 1987; Kim et al., 2005; Sablok et al., 2019), Campanulaceae (Knox et al., 1993; Cosner et al., 2004; Knox, 2014; Knox and Li, 2017; Uribe-Convers et al., 2017), Fabaceae (Kolodner and Tewari, 1979; Palmer and Thompson, 1981; Lavin et al., 1990; Doyle et al., 1996; Cai et al., 2008; Martin et al., 2014; Schwarz et al., 2015; Wang et al., 2017), Geraniaceae (Palmer et al., 1987; Chumley et al., 2006; Guisinger et al., 2011; Weng et al., 2014; Roschenbleck et al., 2017; Weng et al., 2017), Oleaceae (Lee et al., 2007), Plantaginaceae (Zhu et al., 2016; Kwon et al., 2019; Asaf et al., 2020), and Poaceae (Palmer and Thompson, 1982; Doyle et al., 1992; Michelangeli et al., 2003; Burke et al., 2016; Liu et al., 2020).

In the family Papaveraceae, *Corydalis* belongs to the Fumarioideae subfamily. It comprises more than 465 species and is the largest genus in the Papaveraceae family (Zhang et al., 2008). Some *Corydalis* species have medicinal properties and have tremendous potential against hepatitis, tumor, muscular pain, and cardiovascular diseases (Luo et al., 1984; Zhang et al., 2016). The morphological characteristics of the *Corydalis* species are diversified and adapt to various habitats, such as grasslands, forests, riversides, shrubs, and cliffs. In addition, the *Corydalis* species can grow from sea level to more than 6,000 meters in elevation, which is of great interest to ecologists and evolutionary biologists (Niu et al., 2014; Niu et al., 2017). Moreover, these species are distributed widely from the north temperate regions, specifically Qinghai–Tibet Plateau (QTP), to southeast China, Myanmar, the Korean peninsula, and Japan. Xu and Wang (2021) reported that the *Corydalis* species had undergone severe and rapid differentiation (Xu and Wang, 2021). The *Corydalis* plastomes must have also experienced a sequence of genetic shifts to adapt to the radically altered environment. Therefore, ecology biologists are interested in understanding how the plastome structure and its content have fluctuated on a fine scale in the evolutionary period and when rare plastome rearrangements were derived, which is why these modifications occurred (Xu and Wang, 2021). Nevertheless, few *Corydalis* species have been reported since 2019, and a large-scale plastome rearrangement in their structure has been observed. Thus far, twenty *Corydalis* cp genomes have been sequenced, but most studies have reported only briefly. Among these, two research articles explained the genomes in detail. Xu and Wang (2021) used six, and Ren et al. (2021) used two *Corydalis* plastomes for comparative studies (Ren et al., 2021; Xu and Wang, 2021). On the other hand, there are no extensive comparative studies of *Corydalis* plastomes to understand their genome rearrangement patterns, such as

inversion, relocation, expansion, and contraction of IR regions, and their molecular evolution patterns in detail. In addition, divergent time-related molecular studies of the *Corydalis* species could not be found. Therefore, a new plastome of *C. platycarpa* was sequenced, and detailed comparative genomic analyses of all the publicly available twenty *Corydalis* plastomes were conducted. Based on this, this study examined the complexity of the genome structure and rearrangement, gene content, gain/loss of genes and introns, repeats, RNA editing, nucleotide diversity, and adaptive evolution of the *Corydalis* plastomes. Furthermore, the phylogenetic position and divergent time of the *Corydalis* lineages were estimated.

Materials and methods

Genomic DNA isolation and *Corydalis* genome sequencing

Fresh leaves of *C. platycarpa* were sampled from Cheongok mountain, Bonghwa-gun, South Korea (geospatial coordinates: N37°4′9″, E128°57′47″). A voucher specimen (YNUH22C183) was deposited at Yeungnam University Plant Herbarium, Gyeongsan, South Korea. The total gDNA was extracted from the fresh *Corydalis* leaves by a modified CTAB method (Doyle, 1990). Next-generation sequencing was performed with an Illumina HiSeq2500 by Phyzen Ltd., South Korea. The paired-end (PE) library (2 × 150 bp) was constructed using TruSeq PCR free kit and then paired reads with 550 bp insert size were sequenced and ~3 GB of raw data were obtained. FastQC v0.11 (Andrews, 2010) was used to check the low-quality reads, which were removed using Trimmomatic 0.39 (Bolger et al., 2014).

Assembly and annotation of the *Corydalis* chloroplast genome

For the *de novo* chloroplast (cp) genome assembly, the plastid-like reads were obtained from clean reads using the GetOrganelle pipeline v1.7.6.1 (Jin et al., 2020b). The filtered reads were then assembled using SPAdes v3.15.2 (Nurk et al., 2013) for the circular cp genome assembly in the paired-end mode. The assembled *C. platycarpa* cp genome coverage is 26,922×. The complete cp genome sequence and gene annotation were made using the online DOGMA program (Wyman et al., 2004) along with the cp genome annotations of *Nicotiana tabacum* (NCBI Reference sequence: NC_001879). Manual curation was carried out to adjust the start and stop codons of protein and ribosomal coding genes. The *Corydalis* cp genome circular map was drawn using OGDRAW v1.3.1 (Lohse et al., 2007). The annotated genome sequence was submitted to GenBank and assigned the accession number OP142703.

Chloroplast genome sequence divergence and comparison

The newly sequenced *C. platycarpa* cp genome and other 20 publicly available *Corydalis* plastomes (Supplementary Table S1) were compared to determine the cp genome structure synteny and identify the possible rearrangements with the plastome of *N. tabacum* as a reference using Mauve v1.1.3 with the progressiveMauve algorithm (Katoh et al., 2019). A single IR region was used in this analysis. The schematic diagram was drawn manually based on their plastome structure to access the expansion/contraction of the LSC, IR, and SSC junctions of the 21 *Corydalis* plastomes. The entire plastome sequences of all the *Corydalis* were used to visualize the sequence similarity using mVISTA in Shuffle-LAGAN mode (Frazer et al., 2004), with the default parameters and *C. platycarpa* plastome used as a reference.

Analyses of repetitive sequences

The simple sequence repeats (SSR) motifs were analyzed in the 21 plastomes of *Corydalis* using MISA v2.1 (Thiel et al., 2003) with the smallest number of repeats set to ten repetitions for mononucleotide SSRs, six repeat units for dinucleotide SSRs, and five repeat units for tri, tetra, penta, and hexanucleotide SSRs. The tandem repeats were searched using the Phobos Tandem Repeats Finder v1.0.6 with the parameters 1 for the match, −5 for mismatch and gap, and 0 for N positions (Mock et al., 2017). In addition, the forward, reverse, complement, and palindromic repeats were detected using REPuter with a Hamming distance of 3, 90% minimum sequence identity, and 30 bp of a minimal repeat size (Kurtz et al., 2001). For all these analyses, one copy of the IR region was used.

Analyses of the genetic divergence

All 59 protein-coding genes were extracted and aligned individually using Geneious Prime (Biomatters, New Zealand) to evaluate the genetic divergence of the 21 *Corydalis* plastomes. All gaps and missing data were excluded before the analysis. The genetic divergence of 21 *Corydalis* plastomes was calculated by applying nucleotide diversity (π) and the total number of polymorphic sites in the DnaSP v6.12.03 (Librado and Rozas, 2009).

Analysis of RNA editing sites in the protein-coding genes

The predictive RNA Editor for Plants (PREP) suite was applied to analyze the potential RNA editing sites in the protein-

coding genes of the 21 *Corydalis* plastomes. The PREP-cp program has 35 reference genes explaining the RNA editing sites in the cp genomes (Mower, 2009). Therefore, 35 protein-coding genes of the *Corydalis* plastomes were utilized. In the present analysis, the cut-off value was set to 0.8.

Analysis of substitution rate

The complete cp genome of *C. platycarpa* was compared with the other 20 *Corydalis* plastomes. The synonymous (K_S) and non-synonymous (K_A) substitution rates were analyzed by extracting the identical specific 59 functional protein-coding DNA sequences and translating them into protein sequences and aligning them independently using Geneious Prime (Biomatters, New Zealand). The synonymous and non-synonymous substitution rates were assessed in DnaSP v6.12.03 (Librado and Rozas, 2009). Similarly, the substitution analyses for all the 37 Ranunculales and all the Ranunculales (16 taxa) excluding the genus *Corydalis* cp genomes were compared.

Analysis of positive selection

Positive selection analysis was performed based on substitution rate analyses of the 21 *Corydalis* plastomes. The site-specific model was employed using EasyCodeML v1.4 (Gao et al., 2019) to investigate the positive selection analysis. The 24 protein-coding gene sequence was aligned individually using the MAFFT alignment v1.5 (Katoh et al., 2019), and the maximum likelihood phylogenetic tree was built using RAxML v. 7.2.6 (Stamatakis et al., 2008). The codon substitution models, likelihood ratio test and the Bayes Empirical Bayes (BEB) analysis were conducted as described earlier.

Analysis of phylogenetic tree

Thirty-seven cp genomes from the order Ranunculales were selected to construct a phylogenetic tree with *N. tabacum* selected as an outgroup, determine the location of the *C. platycarpa* in the order Ranunculales, and analyze the phylogenetic correlation of the *Corydalis* genus. The cp genome sequences of 29 species across the Papaveraceae family corresponding to two subfamilies (Fumarioideae and Papaveroideae) were downloaded. In addition, two chloroplast genomes of each subfamily of Berberidaceae, Ranunculaceae, Menispermaceae, and Ciraeasteraceae were included in this analysis (Supplementary Table S1). The 59 protein-coding genes shared by 38 plastomes were concatenated, aligned and saved in PHYLIP format using Clustal X v2.1 (Larkin et al., 2007). The Maximum-Likelihood (ML) tree was built using RAxML v7.2.6 with a General Time Reversible + Proportion Invariant model. One thousand non-parametric bootstrap replicates were performed to estimate the support of the data for each internal branch of the phylogeny (Stamatakis et al., 2008).

Analysis of the evolutionary rate

Molecular divergence analysis of the Ranunculales lineages was performed with the Bayesian inference through Bayesian Markov chain Monte Carlo (MCMC) sampling implemented in BEAST v1.4 (Drummond and Rambaut, 2007) with a few modifications, as described earlier (Raman et al., 2021). A relaxed-clock log-normal model was applied using MCMC (500 million steps, sampled every 1000 generations, burn-in of 10%). A maximum clade credibility (MCC) tree was analyzed using TreeAnnotator v2.1.2 (Center for Computational Evolution, University of Auckland, New Zealand). Multiple calibration points were set for the divergence of the Berberidaceae subfamily, such as *Berberis bealei* at 88.94 mya (71.13–100.39 million years ago (mya), 78.22 mya (62.05–90.8 mya) for *Jeffersonia diphylla*, 62 mya (46.9–75.65 mya) for *Epimedium koreanum* and *Diphylleia cymosa*, 55.79 mya (38.65–73.14 mya) for *Nandina domestica* and 13.68 mya (8.05–21.75 mya) for *Caulophyllum robustum* and *Gymnospermium microrrhynchum*, which were employed with a log-normal distribution (Wang et al., 2016a).

Results

General features of the *Corydalis* chloroplast genome

The complete chloroplast (cp) genome sequence of *Corydalis platycarpa* (GenBank: OP142703) is 192,20 bp, with an inverted repeat (IR) of 42,640 bp separating a large single-copy (LSC) region of 96,492 bp and a small single-copy region of 10,247 bp (Figure 1). The average G+C content of the cp genome was 40.4%. The *C. platycarpa* cp genome includes 112 unique genes, such as 78 protein-coding, 30 tRNA, and four rRNA genes. In 112 genes, nine protein-coding and six tRNA genes contained a single intron, and *ycf3* and *rps12* encoded two introns, whereas *clpP* coded for three introns. Moreover, 26 genes were replicated in IR regions, fourteen involving protein-coding, eight tRNA, and four rRNA genes (Supplementary Table S2). The gene *accD* was entirely lost in the cp genome of *C. platycarpa*. In addition, the *C. platycarpa* was compared with other *Corydalis* species and other Fumarioideae (Table 1) and Papaveroideae plastomes (Figure 2; Supplementary Table S3). The average plastome size of the Fumarioideae was 177 kb, whereas the Papaveroideae was only 156.5 kb (Figure 2A). Similarly, the GC content of Fumarioideae and Papaveroideae was 40.7 and 38.7%,

10.3389/fpls.2022.1043740

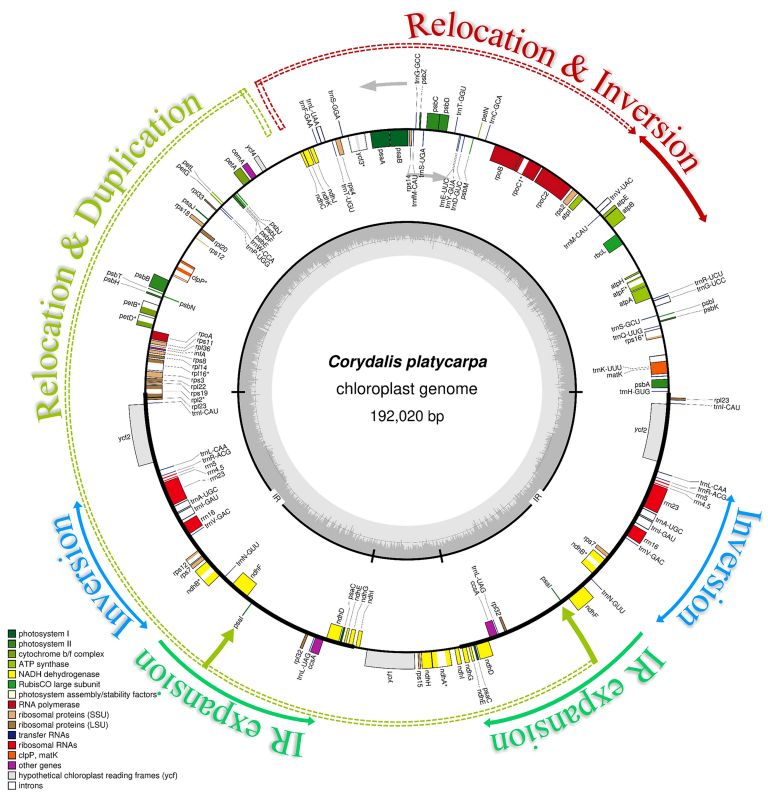


FIGURE 1
Circular chloroplast genome map of *Corydalis platycarpa*. Genes drawn outside the circle are transcribed clockwise, and those inside are counterclockwise. Genes belonging to different functional groups are color-coded. The darker grey in the inner circle shows the GC content, while the lighter grey shows the AT content.

respectively (Figure 2B). In addition, the average length of the LSC region of Fumarioideae and Papaveroideae was 90.3 kb and 85.5 kb, and 10.78 and 18.2 kb of the SSC and 38 and 26.3 kb of the IR regions, respectively (Figures 2C–E).

Comparative analysis of the *Corydalis* chloroplast genome structure

The mauve alignment revealed many rearrangements in the cp genome of the *C. platycarpa* and their relatives of *Corydalis* species (Supplementary Figure S1). Therefore, many events, namely, inversion, translocation, expansion, and contraction duplication, occurred in the SC and IR regions in the cp genomes of *Corydalis*. In the LSC region, the *rps16* gene was relocated within the LSC region of *C. adunca; rbcL – trnV*-UAC inversion and relocation occur in all the cp genomes of *Corydalis* except the species of *C. edulis* and *C. shensiana*. Similarly, the *ndhB – trnR*-ACG inversion was found in all the IR regions of the *Corydalis* cp genomes except the *C. edulis* and *C. shensiana* plastomes. In addition, a *ndhI – ycf1* inversion occurs in the *C. pauciovulata*. In addition, the expansion and contraction of the SC and IR boundaries of the 21 *Corydalis* cp genomes were evaluated using comparative analyses of the genes across the boundary regions (Figure 3). The *rps19* gene straddled the boundary of the LSC/IRB regions of the *C. shensiana*, *C. lupinoides*, and *C. edulis* cp genomes, whereas the *rpl2* gene straddled the LSC/IRB regions of the remaining 18 *Corydalis* cp genomes that lead to the length of LSC regions varies from 82 kb to 98.4 kb (Figure 2). In contrast, the IR regions are highly expanded in most *Corydalis* cp genomes ranging from 22.7 to 52.2 kb. The *ndhF* gene is spanned in the IRB/SSC region in the *C. shensiana* and *C. edulis* cp genomes. Nevertheless, *ndhI, ycf1, rps15, rpl32, trnN*, and *ndhH* genes traversed the remaining *Corydalis* cp genomes due to the relocation, inversion, and expansion of the IR regions. Correspondingly, contraction occurs in the SSC region in most of the *Corydalis* cp genomes (Figures 2, 3) that affect the shuffling of the boundary genes (*ndhA, ndhI, rps15, trnfM, ycf1, trnN*, and *ndhA*) in the SSC/IRA regions. Similarly, most of the *Corydalis* genome encodes the *rpl2* pseudogene in the IRA/LSC boundary regions.

TABLE 1 The basic genomic characteristics of 21 *Corydalis* plastomes.

S. No.	Species name	Genome (bp)					Unique genes				Total genes		
		Total	LSC	IR	SSC	GC (%)	coding (cd)	tRNA (t)	rRNA (r)	Total	Gene Lost/ Pseudogene	Duplicated (cd+t+r)	Total
1	***Corydalis platycarpa***	**192,020**	**96,492**	**42,640**	**10,247**	**40.4**	**78**	**30**	**04**	**112**	**01**	**26 (14 + 8+4)**	**138**
2	*Corydalis adunca*	196,128	92,145	47,226	9,531	41.0	71	30	04	105	08	22 (10 + 8+4)	127
3	*Corydalis conspersa*	187,810	92,280	47,375	780	40.8	67	30	04	101	12	25 (13 + 8+4)	126
4	*Corydalis davidii*	165,416	85,352	39,867	330	40.7	67	30	04	101	12	21 (9 + 8+4)	122
5	*Corydalis edulis*	154,395	81,999	26,250	19,504	40.2	77	29	04	110	03	18 (5 + 7+4)	128
6	*Corydalis fangshanensis*	192,554	98,393	42,263	9,635	40.3	78	30	04	112	01	26 (14 + 8+4)	138
7	*Corydalis filistipes*	169,237	97,016	28,741	14,640	41.2	77	30	04	111	02	17 (6 + 7+4)	128
8	*Corydalis hsiaowutaishanensis*	188,784	88,558	44,070	12,086	40.8	77	30	04	111	02	25 (13 + 8+4)	136
9	*Corydalis impatiens*	197,317	89,790	52,211	3,105	40.7	69	30	04	103	10	23 (11 + 8+4)	126
10	*Corydalis inopinata*	181,335	91,727	44,053	1,502	40.9	68	29	04	101	12	21 (9 + 8+4)	122
11	*Corydalis lupinoides*	178,650	85,220	46,377	1,506	40.8	67	29	04	100	13	19 (7 + 8+4)	119
12	*Corydalis maculata*	165,066	86,472	28,918	20,758	41.0	77	30	04	111	02	17 (5 + 8+4)	128
13	*Corydalis mucronifera*	176,217	85,579	44,778	1,082	40.2	72	30	04	106	07	26 (14 + 8+4)	132
14	*Corydalis namdoensis*	169,818	96,301	29,733	14,051	41.1	77	30	04	111	02	17 (6 + 7+4)	128
15	*Corydalis pauciovulata*	161,773	93,238	22,719	23,097	41.5	66	29	04	99	14	16 (5 + 7+4)	115
16	*Corydalis saxicola*	188,060	94,289	41,969	9,833	40.2	78	30	04	112	01	26 (14 + 8+4)	138
17	*Corydalis shensiana*	155,935	82,369	26,344	20,495	40.6	78	29	04	111	02	15 (4 + 7+4)	126
18	*Corydalis ternata*	170,483	88,722	29,514	22,733	41.2	75	30	04	109	04	17 (6 + 7+4)	126
19	*Corydalis tomentella*	190,198	96,701	41,955	9,636	40.3	77	29	04	110	03	26 (14 + 8+4)	136
20	*Corydalis trisecta*	164,354	91,046	28,345	16,618	41.5	75	29	04	108	05	17 (6 + 7+4)	125
21	*Corydalis turschanivoii*	161,534	89,414	29,143	13,834	40.9	77	30	04	111	02	17 (6 + 7+4)	128

Bold represents the species used in the present study.

Comparative analysis of the repeat sequences in the *Corydalis* cp genomes

The results show that the total number of simple sequence repeats (SSRs) ranges from 19 (*C. ternata*) to 51 (*C. pauciovulata*), and the distribution of SSRs differs among the 21 plastomes of *Corydalis* (Figure 4A). Mononucleotides are the most frequent in the SSRs, distributing 88%, followed by dinucleotide and trinucleotides at 11% and 1%, respectively, in the *Corydalis* plastomes (Figure 4B). Among the mononucleotides, all the cp genomes occupy 96% of A and T type SSRs in their genomes, while most of the species lack dinucleotides, such as AG and CT trinucleotides, namely ATG, ATT, CAA, TTA, and TTG (Supplementary Table S4). Similarly, the distribution of tandem repeats in the *Corydalis* cp genomes ranges from 18 to 71. In addition to SSRs and tandem repeats, 1038 dispersed repeats are identified using REPuter (Figure 4A; Supplementary Table S4). Among the *Corydalis* cp genomes, the forward (76%), palindrome (21%) and reverse (3%) repeats are observed (Figure 4C).

RNA editing site analysis in the *Corydalis* cp genomes

The possible RNA editing sites for 35 protein-coding genes were predicted by the PREP suite in the 21 *Corydalis* cp genomes. One thousand and seventy RNA editing sites were detected in their coding genes (Figure 5A; Supplementary Table S5). The number of editing sites varied from the 46 (*C. mucronifera*) to 57 (*C. adunca*) (Figure 5B; Supplementary Table S5). Among the 35 protein-coding genes of the *Corydalis* plastomes, the *rpoB* gene encoded the highest RNA editing sites (143), followed by *rpoC2* (135), *rpoC1* (122), *atpA* (84), *matK* (80), *rps2* (71), *ycf3* (59), *rpl2* (51), *rpl20* (44), and *petD* (42) (Figure 5A; Supplementary Table S5). In the RNA

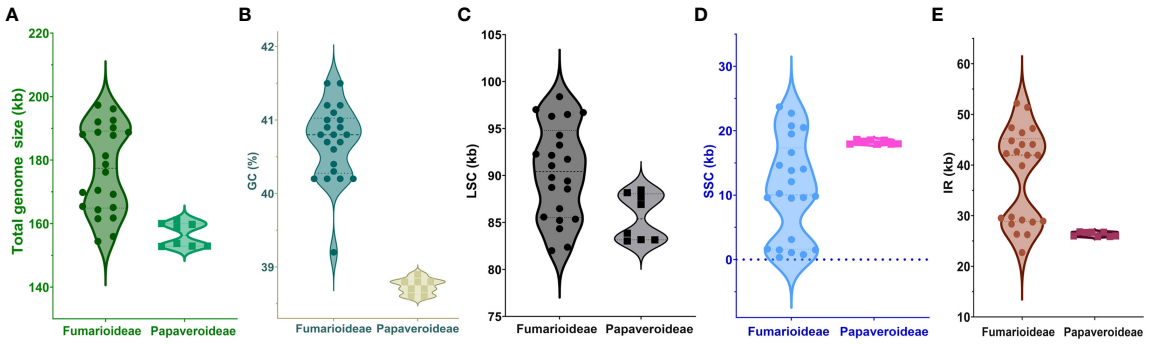


FIGURE 2
Visualization of **(A)** total genome size, **(B)** GC content, **(C)** LSC, **(D)** SSC, and **(E)** IR size of the Fumarioideae plastomes relative to Papaveroideae using a Violin plot.

editing sites, 31% of sites converted serine to leucine, followed by 14% of proline to leucine, 9% of histidine to tyrosine, 8% of proline to serine, 7% of serine to phenylalanine, and 7% of arginine to tryptophane amino acids (Figure 5C; Supplementary Table S5). All predictable RNA editing sites are cytosine to uracil (C–U) transitions, the maximum of which are situated at the second codon position (66%), followed by the first codon position (30%), first and second codon position (4%), besides no transitions at the third codon position (Figure 5D; Supplementary Table S5).

Sequence divergence analysis in the *Corydalis* cp genomes

The sequence divergence of all the 21 plastomes of *Corydalis* was analyzed using mVISTA and sequence identity plots constructed (Figure 6) with the annotated cp genome of *C. platycarpa* as the reference. The results showed that the ribosomal RNA genes in the IR regions were highly conserved and less divergent than the other coding and non-coding sequences in the LSC, SSC, and IR regions. In addition, the nucleotide diversity (Pi) of 59 protein genes in the *Corydalis* cp genomes was calculated. All 59 genes were highly variable regions (>0.03) that are associated with photosynthetic, transcription, and translational processes (Figure 7). Among these 59 genes, *psaC* has the lowest Pi value (0.041), and *rps16* has the highest Pi value 0.642.

Adaptive evolution analysis in the *Corydalis* cp genomes

Fifty-nine shared protein-coding genes of all the 21 *Corydalis* plastomes were used for synonymous (K_S) and non-synonymous (K_A) substitution rates. The results showed that most protein-coding genes have relatively high average K_S values (>0.05) except the *ccsA*, *petN*, *psaJ*, *psbE*, *psbF*, *psbL*, *psbM*, *psbZ*, *rpl32*, *rpl36*, and *rps7* genes (Figure 8A; Supplementary Table S6). In the same way, most of the protein-coding genes are comparatively high average K_A values (>0.02) except *atpA*, *atpB*, *atpI*, *ccsA*, *infA*, *petB*, *petD*, *petG*, *petL*, *petN*, *psaA*, *psaB*, *psaC*, *psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbL*, *psbM*, *psbN*, *psbZ*, *rbcL*, *rpl14*, *rpl16*, *rpl36*, *rps7*, *rps19*, and *ycf3* genes (Figure 8B; Supplementary Table S6). The protein-coding genes, *rps16* and *rps18*, show the highest average K_A/K_S ratio of 1.42 and 1.37, respectively. On the other hand, the K_A/K_S ratios of all the protein-coding genes ranged from 0 to 1.42, with an average ratio of only 0.28 (Figure 8A; Supplementary Table S6). Similarly, all the 37 Ranunculales taxa were analyzed for substitution analysis and revealed that the K_A/K_S ratios of all the protein-coding genes differed from 0 to 6.83, with an average ratio of 0.21 (Supplementary Figure S2; Supplementary Table S7). Furthermore, the substitution analysis of all the 59 protein-coding genes of Ranunculales, excluding the genus *Corydalis* (16 taxa), revealed that the K_A/K_S rate varied from 0 to 0.89, with the average rate of 0.13 (Supplementary Figure S3; Supplementary Table S8).

Suppose the substitution ratio of the specific protein-coding genes among two cp genomes or the whole genomes is > 1.0. In that case, these genes are considered to be under positive selection. Therefore, the K_A/K_S (ω) ratio of 24 protein-coding genes is > 1.0, and they were analyzed for selective pressure events. The ω_2 ratio of 24 protein-coding genes ranges from 1.0 – 107.1382 in the M2a model (Supplementary Table S9). Bayes empirical Bayes (BEB) analysis was employed to assess the position of coherent selective sites in the 24 protein-coding genes utilizing the M7 vs. M8 model and determine that seventeen sites under possibly positive selection in the four protein-coding genes (*rpl20* – 2; *rpl22* – 3; *rpl23* – 1; *rps2* – 2; *rps3* – 1; *rps4* – 3; *rps11* – 1; *rps14* – 2 and *rps18* – 2) with posterior probabilities >0.95 and 21 sites (*ccsA* – 1; *psbJ* – 1; *psbK* – 1; *rpl20* – 1; *rpl22* – 3; *rps3* – 3; *rps4* – 5; *rps7* – 1; *rps8* – 1; *rps11* – 1 and *rps16* – 3) >0.99 (Supplementary Table S9). The positive

10.3389/fpls.2022.1043740

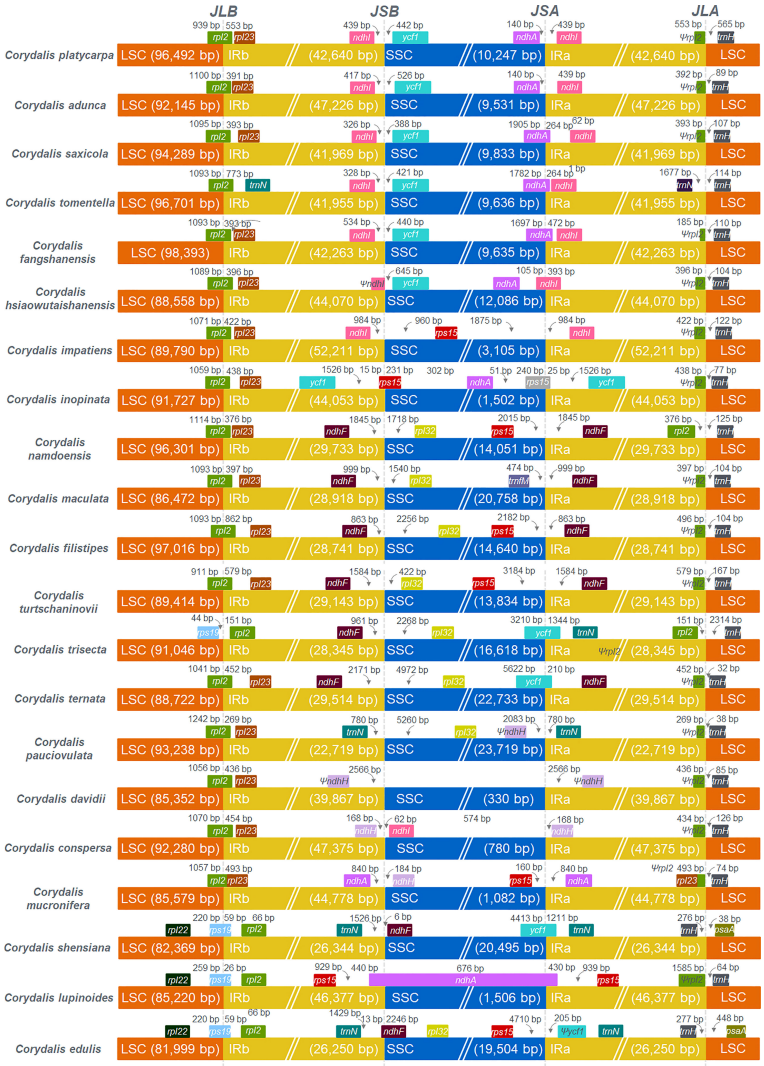


FIGURE 3
Comparison of the borders of LSC, SSC, and IR regions among 21 *Corydalis* chloroplast genomes. JLB indicates the junction line between LSC and IRb; JSB indicates the junction line between SSC and IRb; JSA indicates the junction line between SSC and IRa; JLA indicates the junction between LSC and IRa.

selection models' likelihood ratio test (LRT) statistics against their null models (2ΔLnL) for all 59 genes of 21 *Corydalis* species were evaluated. The 2ΔLnL value ranged from 1.450266 – 115.737378 (Table 2). In contrast, the protein-coding genes *atpB*, *atpE*, *atpF*, *matK*, *psbH*, *psbT*, *rpl16*, *rpl33*, and *rps15* did not positively the encode selected sites in their genes.

Phylogenetic analysis of the Ranunculales

In the present study, 59 concatenated protein-coding genes were used to investigate the phylogenetic relationship of Ranunculales. All the Ranunculales species were clustered into two lineages (clade I and II). In the Papaveraceae lineage, it was grouped into Fumarioideae and Papaveroideae clades. All the *Corydalis* species were clustered into three clades, and *C. adunca* is the basal group in the tree (Figure 9). *C. platycarpa* is the sister to *C. Saxicola*, *C. tomentella*, *C. fangshanensis*, and *C. edulis* and formed one clade. *C. ternata*, *C. turtschaninovii*, *C. flistipes*, *C. maculata*, and *C. namdoensis* formed another clade, whereas *C. davidii*, *C. lupinoides*, *C. pauciovulata*, *C. inopinata*, *C. trisecta*, *C. impatiens*, *C. conspersa*, and *C. mucronifera* formed the third clade. All the *Corydalis* species were supported with strong bootstrap values.

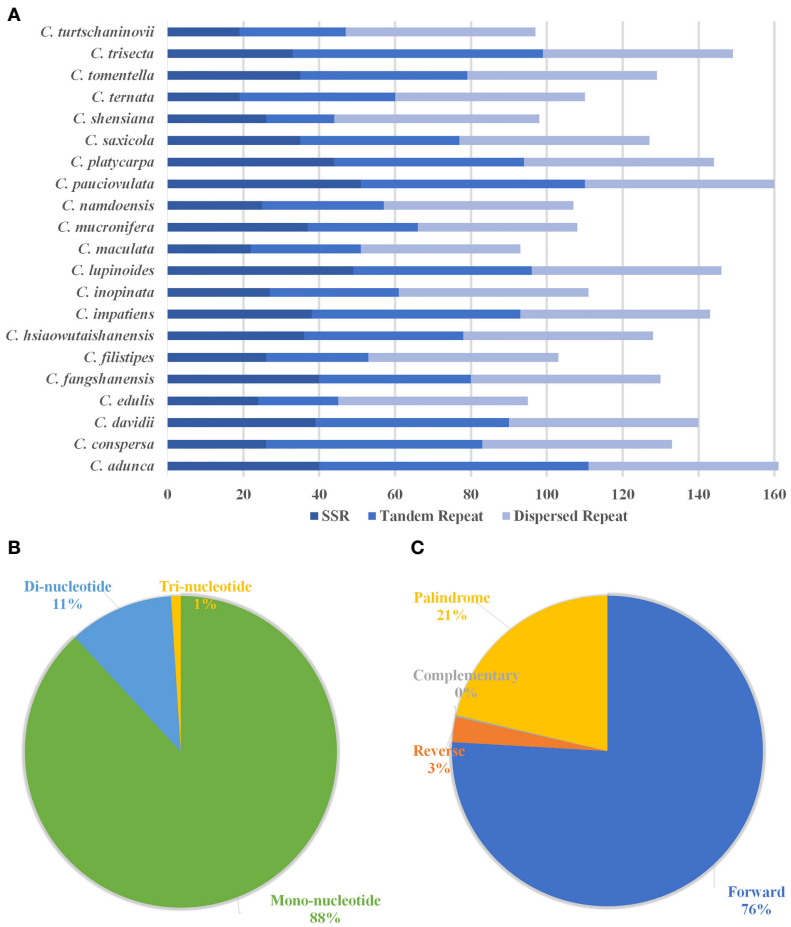


FIGURE 4
Histogram shows the number of repeats in 21 *Corydalis* chloroplast genomes. **(A)** The distribution of simple sequence repeats (SSRs), tandem repeats, and dispersed repeats in the 21 *Corydalis* plastomes. **(B)** Proportion of different SSR repeat types in the 21 plastomes of *Corydalis*. **(C)** The number of different types of dispersed repeats in the 21 *Corydalis* plastomes.

Molecular clock analysis of the Ranunculales

The dataset for 59 protein-coding genes of 37 Ranunculales species was used to estimate the divergent time for the *Corydalis* species. Owing to a lack of calibration points, five other species of Ranunculales were also included. The divergent time was estimated using the previous data of the Ranunculales, which are similar to those obtained in the present study. Among order Ranunculales, the families Circaesteraceae, Menispermaceae, Ranunculaceae, and Berberidaceae diverged 138.13 million years ago (mya) (95% highest posterior density [HPD]: 198.81–90.25 mya). In the Papaveraceae family, Fumarioideae and Papaveroideae diverged 181.08 mya (95% HPD: 272.02–112.44 mya) and 155.65 mya (95% HPD: 223.22–98.61 mya), respectively (Figure 10). The chronogram resulting from a BEAST analysis showed that whole speciation events within *Corydalis* occurred from 98.6 to 1.51 mya. The *C. adunca* diverged from the ancestor of other remaining members of the *Corydalis* species at 98.6 mya (95% HPD: 154.44–56.86 mya). Among the *Corydalis*, the *C. platycarpa* diverged in the early Oligocene period (31.15 mya [95% HPD: 62.15–11.97].

Discussion

Corydalis is the largest genus within the Papaveraceae family and contains more than 465 species (Zhang et al., 2008). Thus far, 20 chloroplast genomes have been sequenced and analyzed. No extensive or comparative studies of the *Corydalis* plastomes have been conducted thus far. Therefore, in the current study, the cp genome of *C. platycarpa* was sequenced and characterized, and comparative studies were carried out with twenty other species of the *Corydalis* genus. The results showed

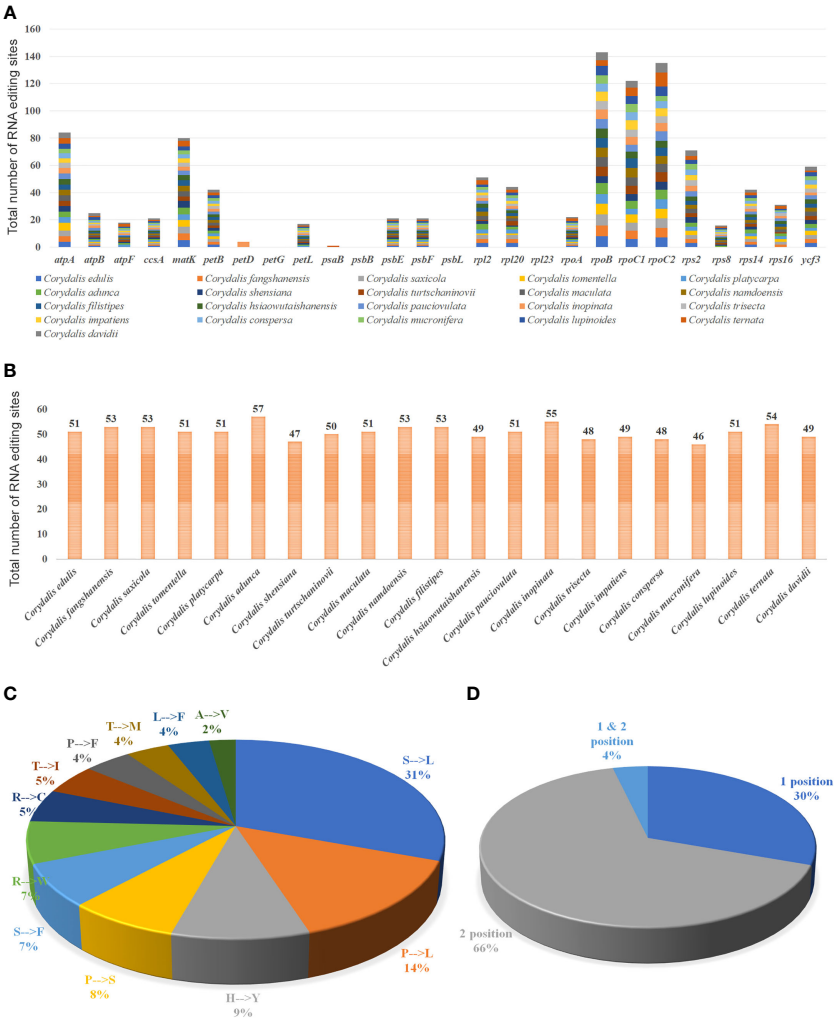


FIGURE 5
Analyses of RNA editing in the 35 protein-coding genes of the 21 *Corydalis* plastomes. **(A)** the distribution of RNA editing sites in the protein-coding genes of each *Corydalis* genome. **(B)** The number of RNA editing sites in each *Corydalis* cp genome. **(C)** Pie diagram represents the conversion percentage of amino acids in the RNA editing sites. **(D)** Represents the RNA editing site in the triplet codon of the nucleotide. S, serine; L, leucine; P, proline; H, histidine; Y, tyrosine; F, phenylalanine; R, arginine; W, tryptophan; C, cysteine; T, threonine; I, isoleucine; M, methionine; A, alanine; V, valine.

that the cp genomes of *Corydalis* displayed the total genome and size of the LSC, SSC, and IR regions (Figure 2; Supplementary Table S3). The gene order and its contents and the GC% content varied and were unusually greater or less than other Papaveraceae cp genomes. The total cp genome size of the *Corydalis* ranged from 154.4 kb (*C. edulis*) to 197.3 kb (*C. impatiens*), and *C. platycarpa* was the fourth largest cp genome (192 kb) in the *Corydalis* and the fifth largest among the Ranunculales cp genomes (Figure 2; Supplementary Table S3). The average size of the *Corydalis* cp genomes is 176.5 kb. In addition, the typical size of the cp genomes of Fumarioideae (*Corydalis* + *Lamprocapnos*) is 177 kb. In contrast, among the Papaveraceae family, the average genome size of the Papaveroideae is only 156.5 kb. Similarly, the average length of the LSC, SSC, and IR of the Fumarioideae was found to be 90.3, 10.78, and 38 kb, respectively. By contrast, the average lengths of the LSC, SSC and IR regions of Papaveroideae were 85.5 kb, 18.2 kb, and 26.3 kb, respectively. The variation in the Furmarioideae was attributed to the expansion of IR regions in their genomes, leading to a shift of the SC genes into the IR regions (Figure 3). In particular, the IR region was extended into the SSC region in most *Corydalis* cp genomes. *C. impatiens* encodes the largest IR region (52.2 kb), and *C. davidii* contains the smallest SSC region (330 bp). This variation in the size of Fumarioideae cp genomes also affects their GC content. The GC content of the *C. pauciovulata* and *C. trisecta* is the highest (41.5%) in the

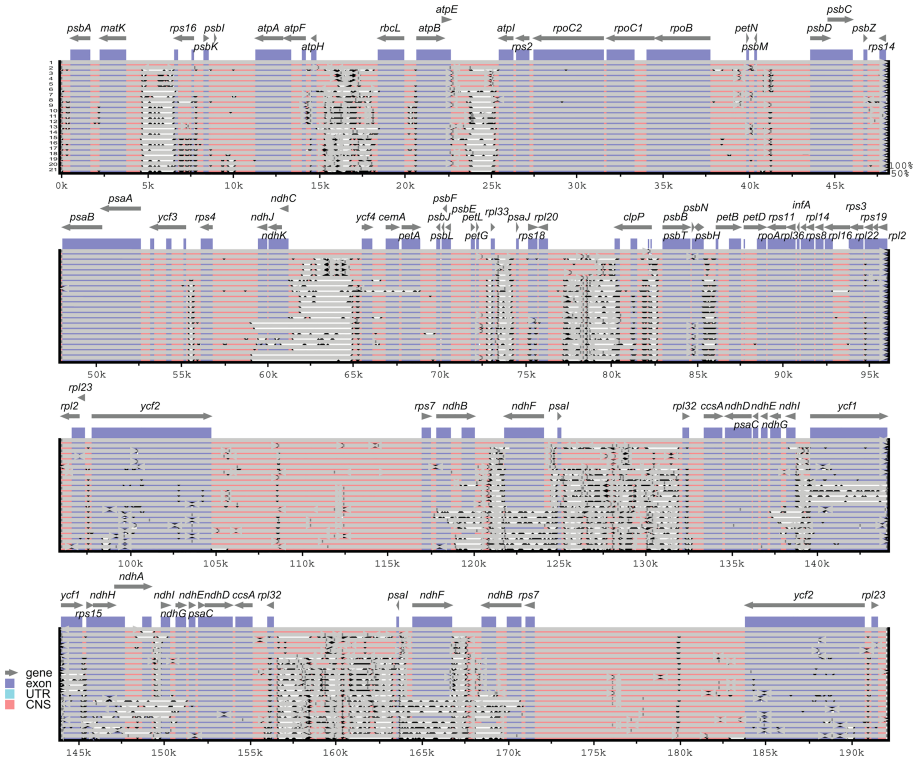


FIGURE 6
mVISTA-based sequence identity plot of 21 *Corydalis* plastomes with *C. platycarpa* as a reference. The gray arrows indicate the direction of the gene transcription. The y-axis represents the percent identity ranging from 50 to 100% is represented by the vertical scale. Coding and non-coding regions are colored purple and pink, respectively. 1. *Corydalis platycarpa*; 2. *Corydalis adunca*; 3. *Corydalis saxicola*; 4. *Corydalis tomentella*; 5. *Corydalis fangshanensis*; 6. *Corydalis hsiaowutaishanensis*; 7. *Corydalis impatiens*; 8. *Corydalis inopinata*; 9. *Corydalis namdoensis*; 10. *Corydalis maculata*; 11. *Corydalis filistipes*; 12. *Corydalis turtschaninovii*; 13. *Corydalis trisecta*; 14. *Corydalis ternata*; 15. *Corydalis pauciovulata*; 16. *Corydalis davidii*; 17. *Corydalis conspersa*; 18. *Corydalis mucronifera*; 19. *Corydalis shensiana*; 20. *Corydalis lupinoides*; 21. *Corydalis edulis*.

Ranunculales plastomes. Commonly, a high GC content imparts more stability to the genome than the AT. In addition, the larger amount of GC base pairs in the genome might impact their adaptation to various adverse environments. On the other hand, the high percentage of AT affects the gene order and its content in the Fumarioideae cp genomes.

In Ranunculales, many relocations, inversions, and rearrangements occurred with all the Fumarioideae plastomes

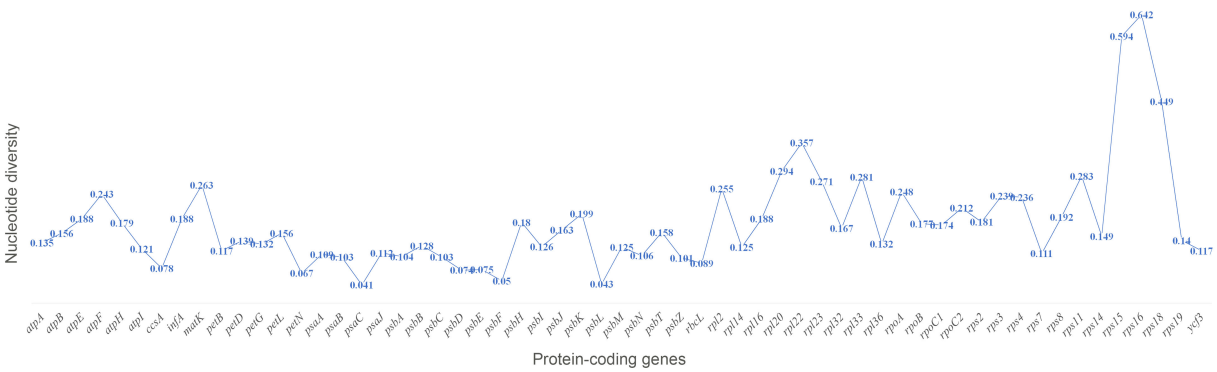


FIGURE 7
Percentage of variable characters (SNPs) in the protein-coding genes of 21 *Corydalis* plastomes.

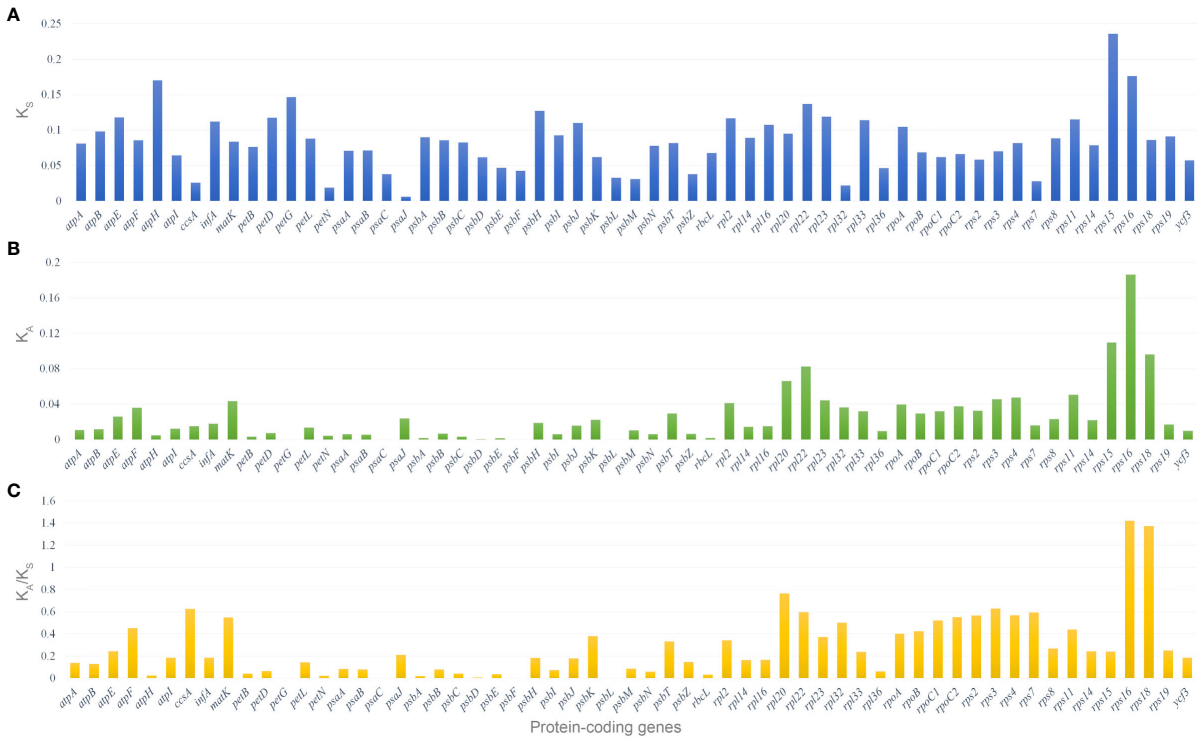


FIGURE 8
Selective pressure of 59 protein-coding genes in the 21 *Corydalis* plastomes. **(A)** K_S, rate of synonymous substitution; **(B)** K_A, rate of non-synonymous substitution; **(C)** K_A/K_S, rate of non-synonymous vs. synonymous substitution.

except for *C. edulis*, *C. shensiana*, and *C. trisecta*. Moreover, at least eight events have occurred in all three LSC, SSC, and IR regions (Figure 9; Supplementary Figure S1) and the following events occurred in the LSC region: (i) a ~ 10 kb of *rbcL* – *trnV*-UAC inverted and relocated into the upstream of *atpH* and downstream of the *atpI* gene in the LSC region (all the Fumarioideae genomes except *Lamprocapnos spectabilis, C. adunca*, *C. edulis*, *C. shensiana* and *C. trisecta*); (ii) the *rps16* gene (LSC) relocated into the IR region (*C. adunca*) and *trnQ* -UUG – *rps16* (LSC) into the IR (*L. spectabilis*); (iii) ~16 kb of *trnG*-GCC – *ndhC* inverted and relocated into the upstream of *psbZ* and downstream of *psbJ* in the LSC region (*C. ternata*); (iv) ~7 kb of *psbK* – *atpH* relocated and ~15 kb of *atpI* - *petN* inverted in the LSC region (*L. spectabilis*) (Park et al., 2018); (v) ~7.5 kb of *trnD*-GUC – *trnfM* relocated into the SSC region (*C. maculata*). The following occurred in the SSC region: (vi) ~10.5 kb of *ndhI* – *ycf1* inverted in the IR region (*C. pauciovulata*): (vii) ~14.4 kb of *ndhB* – *trnR*-ACG inverted (except *C. edulis*, *C. shensiana* and *C. trisecta* in the Fumarioideae); (viii) IR expanded (except *C. edulis*, *C. shensiana*, *C. pauciovulata* and *C. trisecta* in the Fumarioideae). Among these eight events in Fumarioideae, seven events occurred in at least one species of the *Corydalis* cp genome (Figure 9). Furthermore, the translocation and IR expansion analyses were extended in the Ranunculales cp genomes. In the *Berberis bealei* (Berberidaceae) (Ma et al., 2013), the ~13 kb of the LSC region (*rps19* – *psbB*) was transferred to the IR region (Figure 9). Similarly, ~6 kb of the SSC region moved to the IR region, and ~50 kb inversion (*trnQ*-UUG – *accD*) occurred within the LSC region of the *Kingdonia uniflora* (Circaeasteraceae) (Sun et al., 2020) (Figure 9). In contrast, Xu and Wang (2021) reported that the *ndhB* – *trnR*-ACG inversion event occurred in the IR region of the common ancestor of Fumarioideae plastomes (Xu and Wang, 2021). This is due to the limitation of the six plastomes used in their comparative studies. In the present study, 21 *Corydalis* species and one *L. spectabilis* (Park et al., 2018) cp genome (Fumarioideae) were used for comparative analyses, suggesting that the *ndhB* – *trnR*-ACG event did not occur in *C. edulis*, *C. shensiana*, and *C. trisecta* in the Fumarioideae (Figure 9). Therefore, single or all the genome rearrangement/relocation events did not take place in the common ancestor of either the *Corydalis* genera or Fumarioideae clade. Earlier studies reported that several cp genome rearrangements and relocation are feasibly lowest common in angiosperms. To support the present study, previous studies reported similar rearrangement events, such as inversion and relocation of *trnV*-UAC – *rbcL* event in the Oleaceae (Lee et al., 2007) and Campanulaceae (Knox, 2014; Knox and Li, 2017; Uribe-Convers et al., 2017) and *trnQ* -UUG

TABLE 2 Comparison of the likelihood ratio test (LRT) statistics of positive selection models against their null models (2ΔLnL) for across all *Corydalis* species.

Protein-coding genes	Comparison between models	2ΔLnL	*d.f.*	*p*-value
atpB	M0 vs M3	64.72463	4	0
	M1 vs M2a	0	2	1.0
	M7 vs M8	12.88498	2	0.001592437
	M8a vs M8	1.29917	1	0.254364888
atpE	M0 vs M3	8.40558	4	0.077801468
	M1 vs M2a	0	2	1.0
	M7 vs M8	1.450266	2	0.495279515
	M8a vs M8	0.009460	1	0.922517911
atpF	M0 vs M3	18.862202	4	0.000836479
	M1 vs M2a	1.628380	2	0.442998011
	M7 vs M8	2.077926	2	0.353821405
	M8a vs M8	1.641228	1	0.200149937
ccsA	M0 vs M3	33.063826	4	0.000001159
	M1 vs M2a	5.483712	2	0.064450615
	M7 vs M8	5.516278	2	0.063409664
	M8a vs M8	5.483310	1	0.019198871
matK	M0 vs M3	50.242330	4	0
	M1 vs M2a	3.937154	2	0.139655445
	M7 vs M8	5.304254	2	0.070501098
	M8a vs M8	4.425530	1	0.035405116
psbH	M0 vs M3	7.275126	4	0.122043958
	M1 vs M2a	3.999999	2	0.999998000
	M7 vs M8	0.103072	2	0.949769458
	M8a vs M8	0.547220	1	0.459455831
psbJ	M0 vs M3	61.838488	4	0
	M1 vs M2a	58.694068	2	0
	M7 vs M8	59.780422	2	0
	M8a vs M8	59.755664	1	0
psbK	M0 vs M3	64.659550	4	0
	M1 vs M2a	64.659888	2	0
	M7 vs M8	64.735386	2	0
	M8a vs M8	64.773896	1	0
psbT	M0 vs M3	1.969244	4	0.741415908
	M1 vs M2a	0	2	1.0
	M7 vs M8	0.001906	2	0.999047454
	M8a vs M8	0.030365	1	0.861662368
rpl16	M0 vs M3	12.517974	4	0.013887770
	M1 vs M2a	0	2	1.0
	M7 vs M8	2.433942	2	0.296125775
	M8a vs M8	0.191028	1	0.662062396
rpl20	M0 vs M3	82.314682	4	0
	M1 vs M2a	20.167634	2	0.000041750
	M7 vs M8	21.600572	2	0.000020394
	M8a vs M8	21.441600	1	0.000003648
rpl22	M0 vs M3	123.460832	4	0
	M1 vs M2a	26.045432	2	0.00002210
	M7 vs M8	27.377434	2	0.000001135
	M8a vs M8	26.165314	1	0.000000313

(Continued)

10.3389/fpls.2022.1043740

TABLE 2 Continued

Protein-coding genes	Comparison between models	2ΔLnL	*d.f.*	*p*-value
rpl23	M0 vs M3	28.822982	4	0.000008492
	M1 vs M2a	3.272348	2	0.194723632
	M7 vs M8	4.449512	2	0.108093790
	M8a vs M8	3.723888	1	0.053639334
rpl33	M0 vs M3	13.382412	4	0.009550819
	M1 vs M2a	0.117402	2	0.942988681
	M7 vs M8	1.578368	2	0.454215284
	M8a vs M8	0.133798	1	0.714526193
rps2	M0 vs M3	56.521166	4	0
	M1 vs M2a	13.538212	2	0.001148721
	M7 vs M8	14.574584	2	0.000684178
	M8a vs M8	13.592536	1	0.000227087
rps3	M0 vs M3	120.920040	4	0
	M1 vs M2a	35.870314	2	0.000000016
	M7 vs M8	36.737234	2	0.000000011
	M8a vs M8	35.237912	1	0.000000003
rps4	M0 vs M3	164.241576	4	0
	M1 vs M2a	37.675708	2	0.000000007
	M7 vs M8	39.165622	2	0.000000003
	M8a vs M8	38.475640	1	0.000000001
rps7	M0 vs M3	122.425414	4	0
	M1 vs M2a	112.395174	2	0
	M7 vs M8	115.737378	2	0
	M8a vs M8	115.361240	1	0
rps8	M0 vs M3	126.748522	4	0
	M1 vs M2a	75.558924	2	0
	M7 vs M8	76.534240	2	0
	M8a vs M8	71.439114	1	0
rps11	M0 vs M3	56.457308	4	0
	M1 vs M2a	9.114974	2	0.010488383
	M7 vs M8	9.201318	2	0.010045214
	M8a vs M8	9.043680	1	0.002636045
rps14	M0 vs M3	27.682154	4	0.000014466
	M1 vs M2a	3.7647420	2	0.152228743
	M7 vs M8	5.0489140	2	0.080101796
	M8a vs M8	3.764736	1	0.052344121
rps15	M0 vs M3	0	4	1.0
	M1 vs M2a	1.420000	2	0.999929003
	M7 vs M8	1.680000	2	0.999916004
	M8a vs M8	0.025534	1	0.873043652
rps16	M0 vs M3	29.979244	4	0.000004942
	M1 vs M2a	20.030108	2	0.000044722
	M7 vs M8	19.907798	2	0.000047542
	M8a vs M8	19.578680	1	0.000009654
rps18	M0 vs M3	28.119374	4	0.000011797
	M1 vs M2a	9.563052	2	0.008383196
	M7 vs M8	11.694356	2	0.002888038
	M8a vs M8	9.638832	1	0.001905063

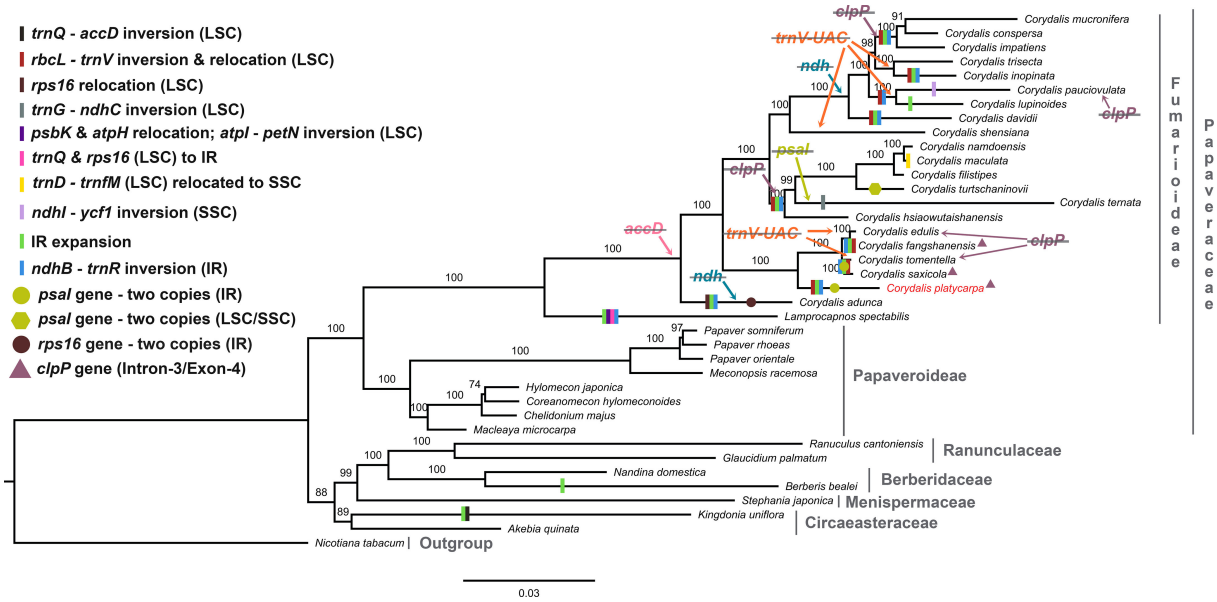


FIGURE 9
Maximum likelihood (ML) tree for 38 taxa based on 59 common plastid protein-coding genes. Values above the branches represent the maximum likelihood bootstrap value.

– *rbcL* in the *Circaeaster agrestis* and *K. uniflora* of Ranunculaceae (Sun et al., 2017; Sun et al., 2020) occurred independently in their plastomes rather than there being a common ancestor.

The gene content in the *Corydalis* plastomes was compared. Usually, the cp genome comprises 79 protein-coding genes (excluding *ycf15* and *ycf68* genes), 30 transfer and four ribosomal RNA genes (Raman et al., 2019; Raman and Park,

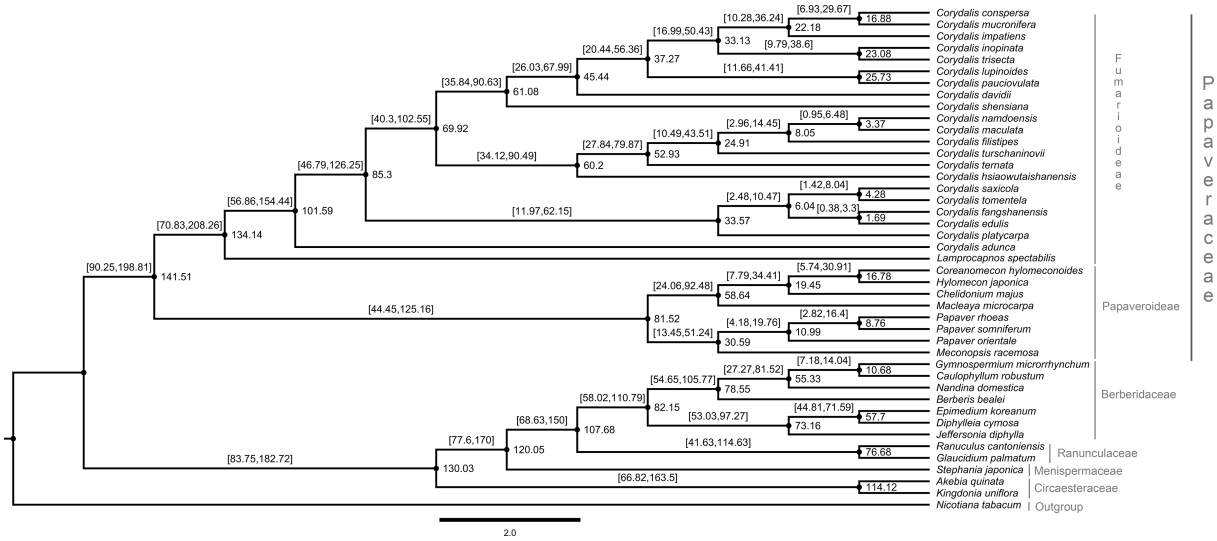


FIGURE 10
Estimation of the divergence time of *Corydalis* species using BEAST based on 59 plastid protein-coding genes of 42 Ranunculales and one outgroup species. The estimated mean ages are shown near the nodes, and blue bars represent 95% high posterior density.

2022). On the other hand, the genus *Corydalis* varied from 66 (*C. pauciovulata*) to 78 protein-coding genes (*C. platycarypa, C. fangshanensis, C. saxicola*, and *C. shensiana*) in their plastomes (Table 1). All the *Corydalis* species lost *accD*, and some of the *Corydalis* lost *clpP*, *ndh*, *rps16*, *psaI*, and *trnV*-UAC genes in their plastomes (Figure 9; Supplementary Table S10). The tRNA and rRNA contents were the same in almost all species except for the loss of the *trnV*-UAC gene in a few *Corydalis* cp genomes (*C. edulis, C. inopinata, C. lupinoides, C. pauciovulata, C. shensiana, C. tomentella*, and *C. trisecta*). In contrast, *trnV*-UAC was reported to be highly conserved in all monocot plants, whereas $tRNA^{Lys}$, $tRNA^{Ala}$, $tRNA^{Ile}$, $tRNA^{Sec}$, $tRNA^{Pyl}$ and suppressor tRNA were absent in most of the monocot cp genomes (Mohanta and Bae, 2017; Mohanta et al., 2019; Mohanta et al., 2020b). By contrast, the absence of all the tRNA genes in the monocots was highly conserved in the *Corydalis* and other closely related cp genomes. Overall, losses of a minimum of one gene to a maximum of fourteen genes occurred in the *Corydalis* plastomes (Table 1; Supplementary Table S10). Among the protein-coding genes, the *accD* gene was lost in all *Corydalis* plastomes. This event happened in the common ancestor of the *Corydalis* lineages (Figure 9). The *accD* encodes one of four subunits of the acetyl-CoA carboxylase enzyme (ACC), which is necessary for fatty acid biosynthesis (Elborough et al., 1996; Sasaki and Nagano, 2004). Moreover, this enzyme is involved in the first process. Kode et al., 2005 suggested that the loss of the *accD* is detrimental to the plants, as observed in a study of tobacco (Kode et al., 2005). Earlier studies confirmed that the missing *accD* gene in the plastome is relocated in the nucleus of angiosperms species, such as *Trifolium repens* (Magee et al., 2010), Campanulaceae (Rousseau-Gueutin et al., 2013), and *Platycodon grandiflorum* (Hong et al., 2017). Because there were no transcriptome data, this study could not confirm whether the cp-encoded *accD* gene was lost wholly or functionally, which was relocated to the nucleus in the genus *Corydalis*. Therefore, further transcriptome studies will be needed to confirm whether the plastid copy is relocated in the *Corydalis* nuclear genome.

Typically, the plastid DNA of most of the higher plants encodes eleven *ndh* genes (Maier et al., 1995; Yukawa et al., 2005) that produce *ndh* polypeptides, forming a thylakoid *ndh* complex (Sazanov et al., 1998; Casano et al., 2000). This *ndh* complex is similar to the mitochondrial complex I, which catalyzes the transfer of electrons from NADH to plastoquinone (Martin and Sabater, 2010). Among the eleven *ndh* genes in the plastomes, the *ndhC*, *ndhK*, and *ndhJ* genes are situated in one transcriptional unit (*ndhC-J* operon) in the LSC region of the plastome (Serrot et al., 2008). The genes *ndhH, ndhA, ndhI, ndhG, ndhE*, and *ndhD* are located in the SSC region (*ndhH-D* operon), which also includes the gene *psaC* (encodes a polypeptide of the photosystem I complex, PSI) between the genes *ndhE* and *ndhD* (Del Campo et al., 2000). In addition, the *ndhF* gene is represented in the SSC region, and two identical copies of the *ndhB* gene exist in IR regions (one on each). The *ndhF* gene and the two *ndhB* genes are possibly transcribed autonomously as monocistronic mRNAs (Martin and Sabater, 2010). In the present study, some *Corydalis* lineages that displayed a wide-ranging pseudogenization or absence of the *ndh* genes in their plastomes were identified (Figure 9; Supplementary Table S10). A similar plastome rearrangement event accompanied by either pseudogenization or a loss of *ndh* genes was also identified in other Ranunculales species, *K. uniflora* (Sun et al., 2017) and Orchidaceae species (Lin et al., 2015). These two events occured independently in these species. In addition, comparative analyses of 2511 cp genomes showed that anyone of the *ndh* gene losses occurred commonly in at least one species of all lineages, such as algae, bryophytes, eudicots, gymnosperms, magnoliids, monocots, protists and pteridophytes (Mohanta et al., 2020a). However, in the *Corydalis* plastomes, at least three to all (eleven genes) *ndh* genes were either pseudogenized or lost in the plastomes of *C. adunca, C. conspersa, C. davidii, C. impatiens, C. inopinata, C. lupinoides, C. mucronifera, C. pauciovulata*, and *C. trisecta plastomes* (Figure 9; Supplementary Table S10). Among the loss of *ndh* genes in the nine plastomes, the *ndhC* and *ndhF* loss occurred in the plastomes of all nine species (Figure 9; Supplementary Table S10). The *C. adunca* is basal for the remaining *Corydalis* species, and the *ndh* gene loss occurred in their genome. The remaining eight species formed a single clade in the phylogenetic tree, and the *ndh* gene loss occurred in this clade, suggesting that after divergence from the *C. shensiana*, it probably occurred in the common ancestor of this clade plastomes (Figure 9). This event appears to be a synapomorphy that occurred at the subgenus level in the *Corydalis* clade. It is probably associated with the rearrangement and relocation of SC and IR genes, boundary shift, and expansion of the IR regions in the *Corydalis* plastomes. Interestingly, all nine species (except *C. pauciovulata*) were distributed predominantly on the Qinghai–Tibet Plateau (QTP) regions. The photosynthetic systems (*ndh* genes) in these plants, might have been lost due to the high altitude conditions, such as low temperatures, strong winds, and low atmospheric pressure, and adapted to their ecological environment.

The *clpP* gene is a proteolytic subunit of the ATP-dependent Clp protease found in higher plant chloroplasts (Shikanai et al., 2001). Usually, the *clpP* encoded three exons spliced by two type II introns in the cp genome (Raman et al., 2019; Raman and Park, 2022). Earlier studies stated that the loss of introns of the *clpP* gene had been determined in the *Geranium*, legume, *Silene*, and *Hypericum* cp genomes (Erixon and Oxelman, 2008; Dugas et al., 2015; Park et al., 2017; Claude et al., 2022). In this study, the *clpP* gene loss also took place in some of the *Corydalis* species, such as *C. conspersa, C. mucronifera, C. impatiens, C. namdoensis, C. maculata, C. filistipes, C.turtschaninovii, C. ternata, C. hsiaowutaishanensis, C. edulis* and *C. tomentalla* (Supplementary Table S10). Among these, *C. conspersa, C. mucnifera*, and *C. impatiens* formed one clade in the phylogenetic tree using 59 protein-coding concatenated datasets and *C. namdoensis, C. maculata, C. filistipes, C.turtschaninovii, C. ternata*, and *C. hsiaowutaishanensis*

formed another (Figure 9). Except for the *C. edulis* and *C. tomentella*, the *clpP* gene loss may have occurred in the common ancestor of these two clades at the subgenera level (Figure 9). On the other hand, it is essential to use additional species to understand the *clpP* loss in the genomes of *Corydalis*. This is supported by a similar type of *clpP* loss that occurred in the common ancestor of the Actinidiaceae family (*Clematoclethra, Actinidia*, and *Saurauia*) (Wang et al., 2016b). In contrast, a few *Corydalis* plastomes (*C. platycarpa, C. saxicola*, and *C. fangshanensis*) encoded four exons and three introns in the *clpP* gene (Supplementary Figure S3). In the *clpP*, there was a ~115 bp insertion after exon 1 in the gene leading to the formation of an additional intron in their plastomes. The inserted nucleotide sequence similarity between the three species was 73.5%. In addition, this insertion sequence in the *clpP* gene was analyzed using BLASTN, but reliable results could not be obtained. The acquisition of one extra intron in the *clpP* gene may be due to the selective pressure in their genome. This could play a role in the evolutionary maintenance of the group II introns and provide more stability to the genome (Petersen et al., 2011). Selective pressures may have been significant and undervalued in the evolution of spliceosomal introns from group II intron progenitors (Chalamcharla et al., 2010). To the best of the authors' knowledge, this rare event has not been identified in any other plastomes. Therefore, further studies will be needed to understand the molecular mechanisms of the *clpP* gene in their plastomes.

In addition, gene duplication also occurred in the *Corydalis* plastomes because of genome rearrangements and boundary shifts. The two copies of the *rps16* gene in the IR regions of *C. adunca* are replaced with *rrn16* in the LSC region, suggesting that at least two rearrangements might have occurred in their plastome simultaneously or independently (Xu and Wang, 2021). At the same time, the *rps16* gene is a pseudogene in the *C. ternata*. Similarly, two copies of the *psaI* gene were found in the IR regions of *C. platycarpa, C. saxicola*, and *C. tomentella*: one copy from LSC and another from SSC in the *C. turtschaninovii* cp genomes (Figure 9; Supplementary Figure S4). Typically the *psaI* gene is located upstream of *accD* and downstream of the *ycf4* gene in the plastomes. The *psaI* gene duplication might have occurred in their cp genomes and was inserted into the IR region. The *psaI* was copied into another IR region due to the copy correction mechanism. In addition, the pseudogenization of the *psaI* gene in the LSC region was identified. A double-strand break might have occurred between the *ndhK* and *psaI* region (which contains *ndhK*, *trnV*-UAC, *trnM*-CAU, *atpE, atpB, rbcL, accD*, and *psaI*). This leads to the excision and inversion of this fragment inserted between the LSC regions in *atpH* and *atpI*. During this process, *accD* gene loss may have occurred in these three species, but this hypothesis could not be concluded for the remaining *Corydalis* plastomes, which might have played a role in the transposon activity. On the other hand, there is no direct evidence of transposable elements with the *Corydalis* cp genome, even though they may have been present transiently.

The SSRs a significant role during genome rearrangements and the recombination process (Ogihara et al., 1988; Milligan et al., 1989; Cole et al., 2018). Therefore, this study analyzed the presence of SSRs in the *Corydalis* plastomes. The distribution of the SSRs in the plastomes of *Corydalis* was quite different from 19 to 51. In addition, the *Corydalis* plastomes distributed many repeats in their genome ranging from 93 to 161 (Figure 4; Supplementary Table S4). Moreover, the presence of repeat sequences does not correlate with their genome rearrangements and relocation events in *Corydalis*. The *C. edulis* has 95 repeat sequences and does not encode major rearrangement events in its genome (Figures 4, 6, 9; Supplementary Table S4). On the other hand, the significant events (inversion, relocation, gene loss, and IR expansion) occurred in the *C. maculata, C. turtschaninovii,* and *C. shensiana* plastomes that encoded similar numbers of repeats (~95) regions in their genomes (Figures 4, 9; Supplementary Table S4). Generally, the RNA editing process arises in the mitochondrial genomes but is less common in the plastomes (Chen et al., 2011; Raman and Park, 2015; Raman et al., 2016). In addition, the seed plant has ~30–40 RNA editing sites in its plastomes (Stern et al., 2010). Nevertheless, all the *Corydalis* have similar numbers (~51) of RNA editing sites in their genomes (Figures 5A–D). This process mainly occurred in the second position, followed by the first position of the triplet codon (Figure 5D). In addition, ~45% of the amino acids were converted to leucine (Figure 5C). Previous studies also reported that C to U RNA editing in the second codon position occurred mainly in plant organelles to enhance the hydrophobic amino acid leucine frequency. Chen et al. (2011) also reported that the closely associated taxa usually contribute to more RNA editing sites due to the evolutionary process but not in this study.

The mVISTA and nucleotide diversity analysis results showed a high degree of variation in both coding and non-coding regions in the *Corydalis* plastomes (Figures 6, 7). The K_A/K_S rate is associated with gene adaptive evolution, such as the positive and purification selection effects (Raman et al., 2020; Raman and Park, 2020). The genes under positive selection might result from natural selection and adaptation to the living environment (Raven et al., 2013; Raman et al., 2020; Raman and Park, 2020; Scobeyeva et al., 2021). Therefore, the substitution rates of all the independent protein-coding genes of 21 *Corydalis* species are averaged. The results showed that the photosynthetic, transcription and transcription-related genes show accelerated non-synonymous rates (Figures 8A, B). Furthermore, the ratio of K_A/K_S (ω) showed that the majority of the protein-coding genes were less than 1, excluding *rps16* and *rps18* genes (Figure 8C). A separate analysis of synonymous and non-synonymous substitution rates was also conducted for all protein-coding genes. Similarly, the substitution analysis of 59 protein-coding genes of all Ranunculales taxa (37 taxa) showed that the K_A/K_S ratio varies from 6.83, with an average ratio of 0.21

(Supplementary Figure S2; Supplementary Table S7). In contrast, the substitution analysis of all the Ranunculales except *Corydalis* taxa (16 taxa) revealed that the K_A/K_S ratio of all these protein-coding genes varies from 0 to 0.89, with an average ratio of 0.13 (Supplementary Figure S3; Supplementary Table S8). This result indicates that all the Ranunculales cp genomes, excluding *Corydalis* taxa, are highly conserved. Therefore, if the ω value is more than 1.0 of the particular protein-coding genes between two plastomes, or the whole genomes of *Corydalis* taxa, these genes are considered to be under positive selection. Therefore, in the present study, 24 protein-coding genes were identified in the *Corydalis* plastomes under positive selection pressure events (Table 2; Supplementary Table S9). In the selective pressure events, six forms of photosynthesis, transcription and translation-related genes were characterized: (i) subunits of ATP synthase (*atpB, atpE*, and *atpF*); (ii) C-type cytochrome synthesis gene (*ccsA*); (iii) maturase (*matK*); (iv) subunits of photosystem II (*psbH, psbJ, psbK*, and *psbT*); (v) large subunits of the ribosome (*rpl16, rpl20, rpl22, rpl23*, and *rpl33*); (vi) small subunit of the ribosome (*rps2, rps3, rps4, rps7, rps8, rps11, rps14, rps15, rps16*, and *rps18*). Among these, fourteen genes (*ccsA, psbJ, psbK, rpl20, rpl22, rpl23, rps2, rps3, rps4, rps8, rps11, rps14, rps16*, and *rps18*) have positively selected sites, providing evidence of the adaptive evolution of proteins (Supplementary Table S9). Genes with various functions, such as genetic and photosynthetic systems, might play a crucial role in the adaptation to the terrestrial ecological environment (Xu et al., 2015; Xu et al., 2020) because most of the *Corydalis* species live at QTP high altitudes and various North, Central, and East Asia terrestrial regions (Supplementary Table S11) and must adapt to high rates of UV radiation, oxygen depletion conditions, temperature fluctuations, and drought stress conditions. Such genes can be a significant genetic foundation for evolutionary adaptation at the chloroplast level (Xu et al., 2020).

The cp genomes are significant genomic resources for reconstructing precise and high-resolution phylogenetic relationships and taxonomic positions in angiosperms (Jansen et al., 2005). In addition to the whole cp genomes, protein-coding genes have been used widely to determine the phylogenetic relationships at every taxonomic level (Li et al., 2017). The phylogenomic analysis showed two distinct clades, such as Papaveraceae and the rest of the Ranunculales. These results are consistent with the previous results. All the *Corydalis* lineages are highly supported with a >97% bootstrap value in the phylogenetic tree, and *C. adunca* is an early divergence species for the remaining *Corydalis* species (Figure 9). No molecular age studies for *Corydalis* species have been reported. Therefore, the divergent times for the genus *Corydalis* were analyzed. The *Corydalis* is estimated to have originated at 98.6 mya (95% HPD: 154.44–56.86 mya) in the early upper Cretaceous period and diverged. It took approximately 16 mya to form the rest of the *Corydalis* species (Figure 10). *C. platycarpa, C. edulis, C. fangshanensis, C. saxicola, C. hsiaowutaishanensis, C. ternata* , *C. turschaninovii, C. filistipes, C. maculata, C. namdoensis*, and *C. shensiana*, distributed in east Asia evolved from 82.86 to 1.51 mya. The remaining eight species are *C. davidii, C. pauciovulata, C. lupoinoides, C. trisecta, C. inopinata, C. impatiens*, and *C. mucronifera* and *C. conspersa*, mainly distributed in the QTP regions. The uplift of the QTP from the period of 25 to 17 mya (Li and Fang, 1999; Wang et al., 2012) changed the environment of East Asia dramatically. The molecular age results of all the eight QTP region *Corydalis* species (44.31 mya [95% HPD: 67.99–26.03 mya] – 15.71 mya [95% HPD: 29.67–6.93 mya]) correlated very well with the uplift of the Qinghai–Tibet Plateau period. This may have caused the radiation of *Corydalis* species during this period. Nevertheless, more taxa will be needed to understand the genome architecture, evolution, and divergence of the *Corydalis* species.

Conclusion

The complete chloroplast genome sequence of *Corydalis platycarpa* species was determined using a *de novo* assembly approach. This is the first comprehensive systematic analysis comparing the plastome rearrangement features and adaptive evolution and inferring phylogenetic and molecular clock relationships using the plastome data of *Corydalis* and its relatives in detail. The comparative analysis showed that Fumarioideae species exhibited high rearrangements, translocation, inversion, duplication, and loss of several protein-coding genes in their genomes. The remaining cp genomes (Papaveroideae, Ranunculaceae, Berberidaceae, Menispermaceae, and Cicaeasteraceae) in the Ranunculales are highly conserved. The *accD* and *ndh* gene loss likely provides a prominent synapomorphic characteristic of the genus *Corydalis*. Phylogenetic and molecular clock studies offer new insights into the systematic relationships between *Corydalis* and will serve as a basis for future research on the phylogenetic, evolution, and biogeography relationships of *Corydalis* species.

Data availability statement

The data presented in the study are deposited in the GenBank repository, accession number OP142703.

Author contributions

GR, SP, and G-HN conceived the project. G-HN provided plant sources. GR designed the experiments. SP and G-HN supervised the project. GR performed the experiments, and analyzed the data, interpreted the results, and wrote and revised the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the National Institute of Biological Resources of Korea (NBR201731201).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1043740/full#supplementary-material

SUPPLEMENTARY TABLE 1
List of taxa and GenBank accession numbers used in the phylogenetic and molecular clock analyses.

SUPPLEMENTARY TABLE 2
List of genes present in the chloroplast genome of *Corydalis platycarpa*.

SUPPLEMENTARY TABLE 3
Summary of the total genome size, GC content, LSC, SSC, and IR regions length, and gene content of 21 *Corydalis* cp genomes.

SUPPLEMENTARY TABLE 4
Distribution of distinct types of repeats in the 21 *Corydalis* cp genomes.

SUPPLEMENTARY TABLE 5
Presence of RNA editing sites, codon position, and amino acid conversion in the protein-coding genes of the 21 *Corydalis* plastomes.

SUPPLEMENTARY TABLE 6
Amount of synonymous and non-synonymous substitution rates present in the 59 protein-coding genes of the 21 *Corydalis* plastomes.

SUPPLEMENTARY TABLE 7
Amount of synonymous and non-synonymous substitution rates present in the 59 protein-coding genes of all the Ranunculales plastomes (37 plastomes).

SUPPLEMENTARY TABLE 8
Amount of synonymous and non-synonymous substitution rates present in the 59 protein-coding genes of all the Ranunculales (16 taxa) except the genus *Corydalis* plastomes.

SUPPLEMENTARY TABLE 9
Comparison of site models, positive selective amino acid loci, and estimation of parameters for 24 protein-coding genes in the *Corydalis* species.

SUPPLEMENTARY TABLE 10
List of pseudogenes and lost genes in the 21 *Corydalis* plastomes.

SUPPLEMENTARY TABLE 11
List of *Corydalis* plants distribution areas.

SUPPLEMENTARY FIGURE 1
MAUVE alignment of Ranunculales plastomes using Geneious Prime. Local collinear blocks are represented by blocks of the same color and linked within each of the alignments.

SUPPLEMENTARY FIGURE 2
Selective pressure analysis for 59 protein-coding genes of all the Ranunculales plastomes (37 taxa). **(A)** K_S: rate of synonymous substitution; **(B)** K_A: rate of non-synonymous substitution; **(C)** K_A/K_S: rate of non-synonymous vs. synonymous substitution.

SUPPLEMENTARY FIGURE 3
Selective pressure analysis for 59 protein-coding genes of all the Ranunculales plastomes (16 taxa) except *Corydalis* taxa. **(A)** K_S: rate of synonymous substitution; **(B)** K_A: rate of non-synonymous substitution; **(C)** K_A/K_S: rate of non-synonymous vs. synonymous substitution.

SUPPLEMENTARY FIGURE 4
Comparison of *clpP* gene in the *Corydalis platycarpa*, *C. fangshanensis*, *C. saxicola* with *C. adunca* plastome.

SUPPLEMENTARY FIGURE 5
Comparison of *psaI* gene in the LSC and IR region of *Corydalis platycarpa* plastome with LSC copy of *C. tomentela psaI*.

References

Andrews, S. (2010). FASTQC. A quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Asaf, S., Khan, A. L., Lubna, K. A., Khan, A., Khan, G., Lee, I. J., et al. (2020). Expanded inverted repeat region with large scale inversion in the first complete plastid genome sequence of *Plantago ovata*. *Sci. Rep.* 10, 3881. doi: 10.1038/s41598-020-60803-y.

Blazier, J. C., Jansen, R. K., Mower, J. P., Govindu, M., Zhang, J., Weng, M. L., et al. (2016). Variable presence of the inverted repeat and plastome stability in *Erodium*. *Ann. Bot.* 117, 1209–1220. doi: 10.1093/aob/mcw065

Bock, R. (2007). "Structure, function, and inheritance of plastid genomes," in *Cell and molecular biology of plastids*. Ed. R. Bock (Berlin, Heidelberg: Springer Berlin Heidelberg), 29–63.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Burke, S. V., Lin, C. S., Wysocki, W. P., Clark, L. G., and Duvall, M. R. (2016). Phylogenomics and plastome evolution of tropical forest grasses (*Leptaspis, streptochaeta*: Poaceae). *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.01993

Cai, Z. Q., Guisinger, M., Kim, H. G., Ruck, E., Blazier, J. C., Mcmurtry, V., et al. (2008). Extensive reorganization of the plastid genome of *Trifolium subterraneum* (Fabaceae) is associated with numerous repeated sequences and novel DNA insertions. *J. Mol. Evol.* 67, 696–704. doi: 10.1007/s00239-008-9180-7

Casano, L. M., Zapata, J. M., Martin, M., and Sabater, B. (2000). Chlororespiration and poising of cyclic electron transport - plastoquinone as electron transporter between thylakoid NADH dehydrogenase and peroxidase. *J. Biol. Chem.* 275, 942–948. doi: 10.1074/jbc.275.2.942

Chalamcharla, V. R., Curcio, M. J., and Belfort, M. (2010). Nuclear expression of a group II intron is consistent with spliceosomal intron ancestry. *Genes Dev.* 24, 827–836. doi: 10.1101/gad.1905010

Chen, H., Deng, L., Jiang, Y., Lu, P., and Yu, J. (2011). RNA Editing sites exist in protein-coding genes in the chloroplast genome of *Cycas taitungensis*. *J. Integr. Plant Biol.* 53, 961–970. doi: 10.1111/j.1744-7909.2011.01082.x

Choi, I. S., Jansen, R., and Ruhlman, T. (2019). Lost and found: Return of the inverted repeat in the legume clade defined by its absence. *Genome Biol. Evol.* 11, 1321–1333. doi: 10.1093/gbe/evz076

Chumley, T. W., Palmer, J. D., Mower, J. P., Fourcade, H. M., Calie, P. J., Boore, J. L., et al. (2006). The complete chloroplast genome sequence of *Pelargonium x hortorum*: Organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. *Mol. Biol. Evol.* 23, 2175–2190. doi: 10.1093/molbev/msl089

Claude, S.-J., Park, S., and Park, S. (2022). Gene loss, genome rearrangement, and accelerated substitution rates in plastid genome of *Hypericum ascyron* (Hypericaceae). *BMC Plant Biol.* 22, 135. doi: 10.1186/s12870-022-03515-x

Cole, L. W., Guo, W., Mower, J. P., and Palmer, J. D. (2018). High and variable rates of repeat-mediated mitochondrial genome rearrangement in a genus of plants. *Mol. Biol. Evol.* 35, 2773–2785. doi: 10.1093/molbev/msy176

Cosner, M. E., Raubeson, L. A., and Jansen, R. K. (2004). Chloroplast DNA rearrangements in Campanulaceae: phylogenetic utility of highly rearranged genomes. *BMC Evolutionary Biol.* 4, 27. doi: 10.1186/1471-2148-4-27

Del Campo, E. M., Sabater, B., and Martin, M. (2000). Transcripts of the *ndhH-D* operon of barley plastids: possible role of unedited site III in splicing of the *ndhA* intron. *Nucleic Acids Res.* 28, 1092–1098. doi: 10.1093/nar/28.5.1092

Doyle, J. (1990). Isolation of plant DNA from fresh tissue. *Focus* 12, 13–15.

Doyle, J. J., Davis, J. I., Soreng, R. J., Garvin, D., and Anderson, M. J. (1992). Chloroplast DNA inversions and the origin of the grass family (Poaceae). *Proc. Natl. Acad. Sci. U.S.A.* 89, 7722–7726. doi: 10.1073/pnas.89.16.7722

Doyle, J. J., Doyle, J. L., Ballenger, J. A., and Palmer, J. D. (1996). The distribution and phylogenetic significance of a 50-kb chloroplast DNA inversion in the flowering plant family Leguminosae. *Mol. Phylogenet Evol.* 5, 429–438. doi: 10.1006/mpev.1996.0038

Drummond, A. J., and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7, 214. doi: 10.1186/1471-2148-7-214

Dugas, D. V., Hernandez, D., Koenen, E. J., Schwarz, E., Straub, S., Hughes, C. E., et al. (2015). Mimosoid legume plastome evolution: IR expansion, tandem repeat expansions, and accelerated rate of evolution in *clpP*. *Sci. Rep.* 5, 16958. doi: 10.1038/srep16958

Elborough, K. M., Winz, R., Deka, R. K., Markham, J. E., White, A. J., Rawsthorne, S., et al. (1996). Biotin carboxyl carrier protein and carboxyltransferase subunits of the multi-subunit form of acetyl-CoA carboxylase from *Brassica napus*: cloning and analysis of expression during oilseed rape embryogenesis. *Biochem. J.* 315 (Pt 1), 103–112. doi: 10.1042/bj3150103

Erixon, P., and Oxelman, B. (2008). Whole-gene positive selection, elevated synonymous substitution rates, duplication, and indel evolution of the chloroplast *clpP1* gene. *PLoS One* 3, e1386. doi: 10.1371/journal.pone.0001386

Frailey, D. C., Chaluvadi, S. R., Vaughn, J. N., Coatney, C. G., and Bennetzen, J. L. (2018). Gene loss and genome rearrangement in the plastids of five Hemiparasites in the family Orobanchaceae. *BMC Plant Biol.* 18, 30. doi: 10.1186/s12870-018-1249-x

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., and Dubchak, I. (2004). VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* 32, W273–W279. doi: 10.1093/nar/gkh458

Gao, F. L., Chen, C. J., Arab, D. A., Du, Z. G., He, Y. H., and Ho, S. Y. W. (2019). EasyCodeML: A visual tool for analysis of selection using CodeML. *Ecol. Evol.* 9, 3891–3898. doi: 10.1002/ece3.5015

Guisinger, M. M., Kuehl, J. V., Boore, J. L., and Jansen, R. K. (2011). Extreme Reconfiguration of Plastid Genomes in the Angiosperm Family Geraniaceae: Rearrangements, Repeats, and Codon Usage (vol 28, pg 583, 2011). *Mol. Biol. Evol.* 28, 1543–1543. doi: 10.1093/molbev/msq229

Hong, C. P., Park, J., Lee, Y., Lee, M., Park, S. G., Uhm, Y., et al. (2017). *accD* nuclear transfer of *Platycodon grandiflorum* and the plastid of early Campanulaceae. *BMC Genomics* 18, 607. doi: 10.1186/s12864-017-4014-x

Jansen, R. K., and Palmer, J. D. (1987). A chloroplast DNA inversion marks an ancient evolutionary split in the sunflower family (Asteraceae). *Proc. Natl. Acad. Sci. U.S.A.* 84, 5818–5822. doi: 10.1073/pnas.84.16.5818

Jansen, R. K., Raubeson, L. A., Boore, J. L., Depamphilis, C. W., Chumley, T. W., Haberle, R. C., et al. (2005). Methods for obtaining and analyzing whole chloroplast genome sequences. *Methods Enzymol.* 395, 348–384. doi: 10.1016/S0076-6879(05)95020-9

Jin, D. M., Wicke, S., Gan, L., Yang, J. B., Jin, J. J., and Yi, T. S. (2020a). The loss of the inverted repeat in the putranjivoid clade of Malpighiales. *Front. Plant Sci.* 11, 942. doi: 10.3389/fpls.2020.00942

Jin, J. J., Yu, W. B., Yang, J. B., Song, Y., Depamphilis, C. W., Yi, T. S., et al. (2020b). GetOrganelle: a fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. *Genome Biol.* 21, 241. doi: 10.1186/s13059-020-02154-5

Katoh, K., Rozewicki, J., and Yamada, K. D. (2019). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform.* 20, 1160–1166. doi: 10.1093/bib/bbx108

Kim, K. J., Choi, K. S., and Jansen, R. K. (2005). Two chloroplast DNA inversions originated simultaneously during the early evolution of the sunflower family (Asteraceae). *Mol. Biol. Evol.* 22, 1783–1792. doi: 10.1093/molbev/msi174

Knox, E. B. (2014). The dynamic history of plastid genomes in the *Campanulaceae sensu lato* is unique among angiosperms. *Proc. Natl. Acad. Sci. United States America* 111, 11097–11102. doi: 10.1073/pnas.140336311

Knox, E., Downie, S., and Palmer, J. (1993). Chloroplast genome rearrangements and the evolution of giant lobelias from herbaceous ancestors. *Mol. Biol. Evol.* 10, 414–414. doi: 10.1093/oxfordjournals.molbev.a040017

Knox, E. B., and Li, C. J. (2017). The East Asian origin of the giant lobelias. *Am. J. Bot.* 104, 924–938. doi: 10.3732/ajb.1700025

Kode, V., Mudd, E. A., Iamtham, S., and Day, A. (2005). The tobacco plastid *accD* gene is essential and is required for leaf development. *Plant J.* 44, 237–244. doi: 10.1111/j.1365-313X.2005.02533.x

Kolodner, R., and Tewari, K. K. (1979). Inverted repeats in chloroplast DNA from higher plants. *Proc. Natl. Acad. Sci. U.S.A.* 76, 41–45. doi: 10.1073/pnas.76.1.41

Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* 29, 4633–4642. doi: 10.1093/nar/29.22.4633

Kwon, W., Kim, Y., Park, C. H., and Park, J. (2019). The complete chloroplast genome sequence of traditional medical herb, *Plantago depressa* willd. (Plantaginaceae). *Mitochondrial DNA Part B-Resources* 4, 437–438. doi: 10.1080/23802359.2018.1553530

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., Mcgettigan, P. A., Mcwilliam, H., et al. (2007). Clustal W and clustal X version 2.0. *Bioinformatics* 23, 2947–2948. doi: 10.1093/bioinformatics/btm404

Lavin, M., Doyle, J. J., and Palmer, J. D. (1990). Evolutionary significance of the loss of the chloroplast-DNA inverted repeat in the *Leguminosae* subfamily *Papilionoideae*. *Evolution* 44, 390–402. doi: 10.1111/j.1558-5646.1990.tb05207.x

Lee, J., Cho, C. H., Park, S. I., Choi, J. W., Song, H. S., West, J. A., et al. (2016). Parallel evolution of highly conserved plastid genome architecture in red seaweeds and seed plants. *BMC Biol.* 14, 75. doi: 10.1186/s12915-016-0299-5

Lee, H. L., Jansen, R. K., Chumley, T. W., and Kim, K. J. (2007). Gene relocations within chloroplast genomes of *Jasminum* and *Menodora* (Oleaceae) are due to multiple, overlapping inversions. *Mol. Biol. Evol.* 24, 1161–1180. doi: 10.1093/molbev/msm036

Librado, P., and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25, 1451–1452. doi: 10.1093/bioinformatics/btp187

Li, J., and Fang, X. (1999). Uplift of the Tibetan plateau and environmental changes. *Chin. Sci. Bull.* 44, 2117–2124. doi: 10.1007/BF03182692

Lin, C. S., Chen, J. J., Huang, Y. T., Chan, M. T., Daniell, H., Chang, W. J., et al. (2015). The location and translocation of *ndh* genes of chloroplast origin in the *Orchidaceae* family. *Sci. Rep.* 5, 9040. doi: 10.1038/srep09040

Liu, Q., Li, X., Li, M., Xu, W., Schwarzacher, T., and Heslop-Harrison, J. S. (2020). Comparative chloroplast genome analyses of *Avena*: insights into evolutionary dynamics and phylogeny. *BMC Plant Biol.* 20, 406. doi: 10.1186/s12870-020-02621-y

Li, Y., Zhou, J. G., Chen, X. L., Cui, Y. X., Xu, Z. C., Li, Y. H., et al. (2017). Gene losses and partial deletion of small single-copy regions of the chloroplast genomes of two *Hemiparasitic Taxillus* species. *Sci. Rep.* 7, 12834. doi: 10.1038/s41598-017-13401-4

Lohse, M., Drechsel, O., and Bock, R. (2007). OrganellarGenomeDRAW (OGDRAW): A tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Curr Genet* 52, 267–274.

Luo, D. S., Feng, C. H., and Xia, G. C. (1984). The resources of the Tibetan drugs in Qinghai-Xizang plateau — preliminary studies on the plants of *Corydalis*. *Zhong Cao Yao* 15, 33–36.

Magee, A. M., Aspinall, S., Rice, D. W., Cusack, B. P., Semon, M., Perry, A. S., et al. (2010). Localized hypermutation and associated gene losses in legume chloroplast genomes. *Genome Res.* 20, 1700–1710. doi: 10.1101/gr.111955.110

Maier, R. M., Neckermann, K., Igloi, G. L., and Kossel, H. (1995). Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. *J. Mol. Biol.* 251, 614–628. doi: 10.1006/jmbi.1995.0460

Maliga, P. (2014). *Chloroplast biotechnology : methods and protocols* (New York: Humana Press).

Martin, G. E., Rousseau-Gueutin, M., Cordonnier, S., Lima, O., Michon-Coudouel, S., Naquin, D., et al. (2014). The first complete chloroplast genome of the Genistoid legume *Lupinus luteus*: evidence for a novel major lineage-specific rearrangement and new insights regarding plastome evolution in the legume family. *Ann. Bot.* 113, 1197–1210. doi: 10.1093/aob/mcu050

Martin, M., and Sabater, B. (2010). Plastid *ndh* genes in plant evolution. *Plant Physiol. Biochem.* 48, 636–645. doi: 10.1016/j.plaphy.2010.04.009

Ma, J., Yang, B., Zhu, W., Sun, L., Tian, J., and Wang, X. (2013). The complete chloroplast genome sequence of *Mahonia bealei* (Berberidaceae) reveals a significant expansion of the inverted repeat and phylogenetic relationship with other angiosperms. *Gene* 528, 120–131. doi: 10.1016/j.gene.2013.07.037

Michelangeli, F. A., Davis, J. I., and Stevenson, D. W. (2003). Phylogenetic relationships among Poaceae and related families as inferred from morphology, inversions in the plastid genome, and sequence data from the mitochondrial and plastid genomes. *Am. J. Bot.* 90, 93–106. doi: 10.3732/ajb.90.1.93

Milligan, B. G., Hampton, J. N., and Palmer, J. D. (1989). Dispersed repeats and structural reorganization in subclover chloroplast DNA. *Mol. Biol. Evol.* 6, 355–368. doi: 10.1093/oxfordjournals.molbev.a040558

Mock, T., Otillar, R. P., Strauss, J., Mcmullan, M., Paajanen, P., Schmutz, J., et al. (2017). Evolutionary genomics of the cold-adapted diatom *Fragilariopsis cylindrus*. *Nature* 541, 536–540. doi: 10.1038/nature20803

Mohanta, T. K., and Bae, H. (2017). Analyses of genomic tRNA reveal presence of novel tRNAs in *Oryza sativa*. *Front. Genet.* 8, 90. doi: 10.3389/fgene.2017.00090

Mohanta, T. K., Khan, A. L., Hashem, A., Abd Allah, E. F., Yadav, D., and Al-Harrasi, A. (2019). Genomic and evolutionary aspects of chloroplast tRNA in monocot plants. *BMC Plant Biol.* 19, 39. doi: 10.1186/s12870-018-1625-6

Mohanta, T. K., Mishra, A. K., Khan, A., Hashem, A., Abd Allah, E. F., and Al-Harrasi, A. (2020a). Gene loss and evolution of the plastome. *Genes* 11, 1133. doi: 10.21203/rs.2.16576/v2

Mohanta, T. K., Yadav, D., Khan, A., Hashem, A., Abd Allah, E. F., and Al-Harrasi, A. (2020b). Analysis of genomic tRNA revealed presence of novel genomic features in cyanobacterial tRNA. *Saudi J. Biol. Sci.* 27, 124–133. doi: 10.1016/j.sjbs.2019.06.004

Mower, J. P. (2009). The PREP suite: predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. *Nucleic Acids Res.* 37, W253–W259. doi: 10.1093/nar/gkp337

Mower, J. P., and Vickrey, T. L. (2018). Structural diversity among plastid genomes of land plants. *Plastid Genome Evol.* 85, 263–292. doi: 10.1016/bs.abr.2017.11.013

Niu, Y., Chen, G., Peng, D. L., Song, B., Yang, Y., Li, Z. M., et al. (2014). Grey leaves in an alpine plant: a cryptic colouration to avoid attack? *New Phytol.* 203, 953–963. doi: 10.1111/nph.12834

Niu, Y., Chen, Z., Stevens, M., and Sun, H. (2017). Divergence in cryptic leaf colour provides local camouflage in an alpine plant. *Proc. Biol. Sci.* 284, 20171654. doi: 10.1098/rspb.2017.1654

Nurk, S., Bankevich, A., Antipov, D., Gurevich, A. A., Korobeynikov, A., Lapidus, A., et al. (2013). Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J. Comput. Biol.* 20, 714–737. doi: 10.1089/cmb.2013.0084

Ogihara, Y., Terachi, T., and Sasakuma, T. (1988). Intramolecular recombination of chloroplast genome mediated by short direct-repeat sequences in wheat species. *Proc. Natl. Acad. Sci. U.S.A.* 85, 8573–8577. doi: 10.1073/pnas.85.22.8573

Palmer, J. D. (1983). Chloroplast DNA exists in two orientations. *Nature* 301, 92–93. doi: 10.1038/301092a0

Palmer, J. D. (1985). Comparative organization of chloroplast genomes. *Annu. Rev. Genet.* 19, 325–354. doi: 10.1146/annurev.ge.19.120185.001545

Palmer, J. D., Nugent, J. M., and Herbon, L. A. (1987). Unusual structure of *Geranium* chloroplast DNA: A triple-sized inverted repeat, extensive gene duplications, multiple inversions, and two repeat families. *Proc. Natl. Acad. Sci. U. S. A.* 84, 769–773. doi: 10.1073/pnas.84.3.769

Palmer, J. D., and Thompson, W. F. (1981). Rearrangements in the chloroplast genomes of mung bean and pea. *Proc. Natl. Acad. Sci. U.S.A.* 78, 5533–5537. doi: 10.1073/pnas.78.9.5533

Palmer, J. D., and Thompson, W. F. (1982). Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. *Cell* 29, 537–550. doi: 10.1016/0092-8674(82)90170-2

Park, S., An, B., and Park, S. (2018). Reconfiguration of the plastid genome in *Lamprocapnos spectabilis*: IR boundary shifting, inversion, and intraspecific variation. *Sci. Rep.* 8, 13568. doi: 10.1038/s41598-018-31938-w

Park, S., Ruhlman, T. A., Weng, M.-L., Hajrah, N. H., Sabir, J. S. M., and Jansen, R. K. (2017). Contrasting patterns of nucleotide substitution rates provide insight into dynamic evolution of plastid and mitochondrial genomes of *Geranium*. *Genome Biol. Evol.* 9, 1766–1780. doi: 10.1093/gbe/evx124

Petersen, K., Schöttler, M. A., Karcher, D., Thiele, W., and Bock, R. (2011). Elimination of a group II intron from a plastid gene causes a mutant phenotype. *Nucleic Acids Res.* 39, 5181–5192. doi: 10.1093/nar/gkr105

Raman, G., Choi, K. S., and Park, S. (2016). Phylogenetic relationships of the fern *Cyrtomium falcatum* (Dryopteridaceae) from dokdo island based on chloroplast genome sequencing. *Genes* 7, 115. doi: 10.3390/genes7120115

Raman, G., Lee, E. M., and Park, S. (2021). Intracellular DNA transfer events restricted to the genus *Convallaria* within the asparagaceae family: Possible mechanisms and potential as genetic markers for biographical studies. *Genomics* 113, 2906–2918. doi: 10.1016/j.ygeno.2021.06.033

Raman, G., and Park, S. (2015). Analysis of the complete chloroplast genome of a medicinal plant, *Dianthus superbus* var. *longicalyncinus*, from a comparative genomics perspective. *PLoS One* 10, e0141329. doi: 10.1371/journal.pone.0141329.

Raman, G., and Park, S. (2020). The complete chloroplast genome sequence of the *Speirantha gardenii*: Comparative and adaptive evolutionary analysis. *Agronomy*. 10, 1405. doi: 10.3390/agronomy10091405

Raman, G., and Park, S. (2022). Structural characterization and comparative analyses of the chloroplast genome of Eastern Asian species *Cardamine occulta* (Asian *C. flexuosa* with.) and other cardamine species. *Front. Biosci. (Landmark Ed)* 27, 124. doi: 10.31083/j.fbl2704124

Raman, G., Park, K. T., Kim, J.-H., and Park, S. (2020). Characteristics of the completed chloroplast genome sequence of *Xanthium spinosum*: comparative analyses, identification of mutational hotspots and phylogenetic implications. *BMC Genomics* 21, 855. doi: 10.1186/s12864-020-07219-0

Raman, G., Park, S., Lee, E. M., and Park, S. (2019). Evidence of mitochondrial DNA in the chloroplast genome of *Convallaria keiskei* and its subsequent evolution in the asparagales. *Sci. Rep.* 9 5028. doi: 10.1038/s41598-019-41377-w

Raven, J. A., Beardall, J., Larkum, A. W., and Sanchez-Baracaldo, P. (2013). Interactions of photosynthesis with genome size and function. *Philos. Trans. R Soc. Lond B Biol. Sci.* 368, 20120264. doi: 10.1098/rstb.2012.0264

Ren, F., Wang, L., Li, Y., Zhuo, W., Xu, Z., Guo, H., et al. (2021). Highly variable chloroplast genome from two endangered *Papaveraceae* lithophytes *Corydalis tomentella* and *Corydalis saxicola*. *Ecol. Evol.* 11, 4158–4171. doi: 10.1002/ece3.7312

Roschenbleck, J., Wicke, S., Weinl, S., Kudla, J., and Muller, K. F. (2017). Genus-wide screening reveals four distinct types of structural plastid genome organization in *Pelargonium* (Geraniaceae). *Genome Biol. Evol.* 9, 64–76. doi: 10.1093/gbe/evw271

Rousseau-Gueutin, M., Huang, X., Higginson, E., Ayliffe, M., Day, A., and Timmis, J. N. (2013). Potential functional replacement of the plastidic acetyl-CoA carboxylase subunit (*accD*) gene by recent transfers to the nucleus in some angiosperm lineages. *Plant Physiol.* 161, 1918–1929. doi: 10.1104/pp.113.214528

Ruhlman, T. A., Zhang, J., Blazier, J. C., Sabir, J. S. M., and Jansen, R. K. (2017). Recombination-dependent replication and gene conversion homogenize repeat sequences and diversify plastid genome structure. *Am. J. Bot.* 104, 559–572. doi: 10.3732/ajb.1600453

Sablok, G., Amiryousefi, A., He, X., Hyvonen, J., and Poczai, P. (2019). Sequencing the plastid genome of giant ragweed (*Ambrosia trifida*, Asteraceae) from a herbarium specimen. *Front. Plant Sci.* 10, 218. doi: 10.3389/fpls.2019.00218

Sanderson, M. J., Copetti, D., Burquez, A., Bustamante, E., Charboneau, J. L. M., Eguiarte, L. E., et al. (2015). Exceptional reduction of the plastid genome of saguaro cactus (*Carnegiea gigantea*): Loss of the *ndh* gene suite and inverted repeat. *Am. J. Bot.* 102, 1115–1127. doi: 10.3732/ajb.1500184

Sasaki, Y., and Nagano, Y. (2004). Plant acetyl-CoA carboxylase: structure, biosynthesis, regulation, and gene manipulation for plant breeding. *Biosci. Biotechnol. Biochem.* 68, 1175–1184. doi: 10.1271/bbb.68.1175

Sazanov, L. A., Burrows, P. A., and Nixon, P. J. (1998). The plastid *ndh* genes code for an NADH-specific dehydrogenase: Isolation of a complex I analogue from pea thylakoid membranes. *Proc. Natl. Acad. Sci. U.S.A.* 95, 1319–1324. doi: 10.1073/pnas.95.3.1319

Schwarz, E. N., Ruhlman, T. A., Sabir, J. S. M., Hajrah, N. H., Alharbi, N. S., Al-Malki, A. L., et al. (2015). Plastid genome sequences of legumes reveal parallel inversions and multiple losses of *rps16* in papilionoids. *J. Systematics Evol.* 53, 458–468. doi: 10.1111/jse.12179

Scobeyeva, V. A., Artyushin, I. V., Krinitsina, A. A., Nikitin, P. A., Antipin, M. I., Kuptsov, S. V., et al. (2021). Gene loss, pseudogenization in plastomes of genus *Allium* (Amaryllidaceae), and putative selection for adaptation to environmental conditions. *Front. Genet.* 12, 674783. doi: 10.3389/fgene.2021.674783

Serrot, P. H., Sabater, B., and Martín, M. (2008). Expression of the *ndhCKJ* operon of barley and editing at the 13th base of the mRNA of the *ndhC* gene. *Biol. Plantarum* 52, 347–350. doi: 10.1007/s10535-008-0071-y

Shikanai, T., Shimizu, K., Ueda, K., Nishimura, Y., Kuroiwa, T., and Hashimoto, T. (2001). The chloroplast *clpP* gene, encoding a proteolytic subunit of ATP-dependent protease, is indispensable for chloroplast development in tobacco. *Plant Cell Physiol.* 42, 264–273. doi: 10.1093/pcp/pce031

Stamatakis, A., Hoover, P., and Rougemont, J. (2008). A rapid bootstrap algorithm for the RAxML web servers. *Syst. Biol.* 57, 758–771. doi: 10.1080/10635150802429642

Stern, D. B., Goldschmidt-Clermont, M., and Hanson, M. R. (2010). Chloroplast RNA metabolism. *Annu. Rev. Plant Biol.* 61, 125–155. doi: 10.1146/annurev-arplant-042809-112242

Sun, Y., Deng, T., Zhang, A., Moore, M. J., Landis, J. B., Lin, N., et al. (2020). Genome sequencing of the endangered *Kingdonia uniflora* (Circaeasteraceae, Ranunculales) reveals potential mechanisms of evolutionary specialization. *iScience* 23, 101124. doi: 10.1016/j.isci.2020.101124

Sun, Y. X., Moore, M. J., Lin, N., Adelalu, K. F., Meng, A. P., Jian, S. G., et al. (2017). Complete plastome sequencing of both living species of circaeasteraceae (*Ranunculales*) reveals unusual rearrangements and the loss of the *ndh* gene family. *BMC Genomics* 18, 592. doi: 10.1186/s12864-017-3956-3

Thiel, T., Michalek, W., Varshney, R., and Graner, A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor. Appl. Genet.* 106, 411–422. doi: 10.1007/s00122-002-1031-0

Uribe-Convers, S., Carlsen, M. M., Lagomarsino, L. P., and Muchhala, N. (2017). Phylogenetic relationships of burmeistera (Campanulaceae: Lobelioideae): Combining whole plastome with targeted loci data in a recent radiation. *Mol. Phylogenet. Evol.* 107, 551–563. doi: 10.1016/j.ympev.2016.12.011

Wang, W. C., Chen, S. Y., and Zhang, X. Z. (2016b). Chloroplast genome evolution in Actinidiaceae: *clpP* loss, heterogenous divergence and phylogenomic practice. *PLoS One* 11, e0162324. doi: 10.1371/journal.pone.0162324

Wang, W., Lin, L., Xiang, X. G., Ortiz, R. D., Liu, Y., Xiang, K. L., et al. (2016a). The rise of angiosperm-dominated herbaceous floras: Insights from *Ranunculaceae*. *Sci. Rep.* 6, 27259. doi: 10.1038/srep27259.

Wang, Y. H., Qu, X. J., Chen, S. Y., Li, D. Z., and Yi, T. S. (2017). Plastomes of mimosoideae: structural and size variation, sequence divergence, and phylogenetic implication. *Tree Genet. Genomes* 13, 41. doi: 10.1007/s11295-017-1124-1.

Wang, Y., Zheng, J., Zhang, W., Li, S., Liu, X., Yang, X., et al. (2012). Cenozoic uplift of the Tibetan Plateau: Evidence from the tectonic–sedimentary evolution of the western Qaidam basin. *Geosci. Front.* 3, 175–187.

Weng, M. L., Blazier, J. C., Govindu, M., and Jansen, R. K. (2014). Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats, and nucleotide substitution rates. *Mol. Biol. Evol.* 31, 645–659. doi: 10.1093/molbev/mst257

Weng, M. L., Ruhlman, T. A., and Jansen, R. K. (2017). Expansion of inverted repeat does not decrease substitution rates in *Pelargonium* plastid genomes. *New Phytol.* 214, 842–851. doi: 10.1111/nph.14375

Wicke, S., Schneeweiss, G. M., Depamphilis, C. W., Muller, K. F., and Quandt, D. (2011). The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Mol. Biol.* 76, 273–297. doi: 10.1007/s11103-011-9762-4

Wyman, S. K., Jansen, R. K., and Boore, J. L. (2004). Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20, 3252–3255. doi: 10.1093/bioinformatics/bth352

Xu, Z., Jiang, Y., and Zhou, G. (2015). Response and adaptation of photosynthesis, respiration, and antioxidant systems to elevated CO_2 with environmental stress in plants. *Front. Plant Sci.* 6, 701. doi: 10.3389/fpls.2015.00701

Xu, X. D., and Wang, D. (2021). Comparative chloroplast genomics of *Corydalis* species (Papaveraceae): Evolutionary perspectives on their unusual Large scale rearrangements. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.600354

Xu, S., Wang, J., Guo, Z., He, Z., and Shi, S. (2020). Genomic convergence in the adaptation to extreme environments. *Plant Commun.* 1, 100117. doi: 10.1016/j.xplc.2020.100117

Yukawa, M., Tsudzuki, T., and Sugiura, M. (2005). The 2005 version of the chloroplast DNA sequence from tobacco (*Nicotiana tabacum*). *Plant Mol. Biol. Rep.* 23, 359–365. doi: 10.1007/BF02788884

Zhang, B., Huang, R., Hua, J., Liang, H., Pan, Y., Dai, L., et al. (2016). Antitumor lignanamides from the aerial parts of *Corydalis saxicola*. *Phytomedicine* 23, 1599–1609. doi: 10.1016/j.phymed.2016.09.006

Zhang, M. L., Su, Z. Y., and Lidén, M. (2008). *Flora of China* (Beijing: Science Press).

Zhu, A. D., Guo, W. H., Gupta, S., Fan, W. S., and Mower, J. P. (2016). Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. *New Phytol.* 209, 1747–1756. doi: 10.1111/nph.13743

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 04 January 2023
DOI 10.3389/fpls.2022.1075353

OPEN ACCESS

EDITED BY
Baozhong Duan,
Dali University, China

REVIEWED BY
Hengchang Wang,
Wuhan Botanical Garden (CAS), China
Gurusamy Raman,
Yeungnam University, South Korea
Lei Pan,
Jianghan University, China

*CORRESPONDENCE
Xueyi Zhu
zhuxueyi90@xmu.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

RECEIVED 20 October 2022
ACCEPTED 08 December 2022
PUBLISHED 04 January 2023

CITATION
Xu X, Shen Y, Zhang Y, Li Q, Wang W, Chen L, Chen G, Ng WL, Islam MN, Punnarak P, Zheng H and Zhu X (2023) A comparison of 25 complete chloroplast genomes between sister mangrove species *Kandelia obovata* and *Kandelia candel* geographically separated by the South China Sea. *Front. Plant Sci.* 13:1075353. doi: 10.3389/fpls.2022.1075353

A comparison of 25 complete chloroplast genomes between sister mangrove species *Kandelia obovata* and *Kandelia candel* geographically separated by the South China Sea

Xiuming Xu[1†], Yingjia Shen[1†], Yuchen Zhang[1], Qianying Li[2], Wenqing Wang[1], Luzhen Chen[1], Guangcheng Chen[3], Wei Lun Ng[4], Md Nazrul Islam[5], Porntep Punnarak[6], Hailei Zheng[1] and Xueyi Zhu[1*]

[1]Key Laboratory of the Ministry of Education for Coastal and Wetland Ecosystems, College of the Environment and Ecology, Xiamen University, Xiamen, China, [2]School of Life Sciences, Xiamen University, Xiamen, China, [3]Third Institute of Oceanography, Ministry of Natural Resources, Xiamen, China, [4]China-ASEAN College of Marine Sciences, Xiamen University Malaysia, Selangor Darul Ehsan, Malaysia, [5]Forestry and Wood Technology Discipline, Khulna University, Khulna, Bangladesh, [6]Aquatic Resources Research Institute, Chulalongkorn University, Bangkok, Thailand

In 2003, *Kandelia obovata* was identified as a new mangrove species differentiated from *Kandelia candel*. However, little is known about their chloroplast (cp) genome differences and their possible ecological significance. In this study, 25 whole cp genomes, with seven samples of *K. candel* from Malaysia, Thailand, and Bangladesh and 18 samples of *K. obovata* from China, were sequenced for comparison. The cp genomes of both species encoded 128 genes, namely 83 protein-coding genes, 37 tRNA genes, and eight rRNA genes, but the cp genome size of *K. obovata* was ~2 kb larger than that of *K. candle* due to the presence of more and longer repeat sequences. Of these, tandem repeats and simple sequence repeats exhibited great differences. Principal component analysis based on indels, and phylogenetic tree analyses constructed with homologous protein genes from the single-copy genes, as well as 38 homologous pair genes among 13 mangrove species, gave strong support to the separation of the two species within the *Kandelia* genus. Homologous genes *ndhD* and *atpA* showed intraspecific consistency and interspecific differences. Molecular dynamics simulations of their corresponding proteins, NAD(P)H dehydrogenase chain 4 (NDH-D) and ATP synthase subunit alpha (ATP-A), predicted them to be significantly different in the functions of photosynthetic electron transport and ATP generation in the two species. These results suggest that the energy requirement was a pivotal

factor in their adaptation to differential environments geographically separated by the South China Sea. Our results also provide clues for future research on their physiological and molecular adaptation mechanisms to light and temperature.

KEYWORDS

mangrove, *Kandelia*, chloroplast genome, gene diversity, protein dynamics simulation, environment adaptation

1 Introduction

Mangroves are woody plant communities distributed in tropical and subtropical intertidal zones that play a vital role in reducing the impact of natural disasters and maintaining the coastal ecological environment (Lin, 1999). *Kandelia*, a typical viviparous mangrove species in the Rhizophoraceae family, is widely distributed from Ganges Delta, Burma, through the South China Sea to southern China, and southern Japan (Tomlinson, 1986; Sheue et al., 2003). *Kandelia obovata* (known as *K. candel* until 2003) is one of the most important and dominant mangrove species naturally distributed along the coastal areas of southeastern China with the widest distribution and the highest latitude, including in Hainan, Guangdong, Guangxi, and Fujian, in addition to Hong Kong, Macao, and Taiwan (Li & Lee, 1997; Lin, 1999). As the strongest cold-resistant true mangrove species, it is not only a high-quality mangrove species that has been artificially planted for ecological restoration of northward coastal wetlands in China (Liao & Zhang, 2014), such as Yueqing in Zhejiang Province, but also an ideal plant material suitable for exploring the relationship between genetic variation and geographical distribution of mangrove species.

The genus *Kandeli*a originated from Malabar, India, and *Kandelia candel* (L.) Druce was named because of its candle-like hypocotyl propagule. *Kandelia candel* was once regarded as the only species in the *Kandelia* genus (Hou, 1958; Juncosa & Tomlinson, 1988). Comparative analysis of the morphological characteristics of leaves and propagules of *Kandelia* populations collected from Brunei, Thailand, and Hong Kong indicated that distinctive differences exist among them (Maxwell, 1995). Naskar & Mandal (1999) further found that the anatomical structures of the *Kandelia* populations distributed in India and Taiwan were also significantly different. Analysis of the chromosome numbers and karyotypes between populations in the two geographical regions of *Kandelia* showed 2n = 38 for the Indian populations and 2n = 36 for the Japanese populations (Yoshioka et al., 1984; Das et al., 1995). Based on the abovementioned differences in the morphology, structure, and chromosome number, Sheue et al. (2003) proposed that the populations distributed in the south of the South China Sea (including in India, Burma, Thailand, and Malaysia) continue to be recognized as *K. candel*, and the populations growing in the north of the South China Sea (including in northern Vietnam, southeastern China, and southern Japan) are recognized as a new species named *K. obovata*, after their obovate leaves. Although the *atpB-rbcL* and *trnL-trnF* fragments of the chloroplast (cp) genome were used as molecular markers in the two geographical populations (Chiang et al., 2001), the detailed features of the whole cp genomes between the two population groups and the functions of differential genes in their adaptation to the respective environments need to be further clarified (Takeuchi et al., 2001; Giang et al., 2006; Geng et al., 2008).

The cp genome has the unique advantages of small haploid size, abundant copy number, relatively conservative gene number and arrangement, lack of recombination, and maternal inheritance (Birky, 2001; Rousseau Gueutin et al., 2015). With the development of next-generation DNA sequencing technologies, the complete cp genome has been widely used for plant identification, phylogeny, and evolution studies. Given that cps are important for interactions between a plant species and its environment (including responses to cold, heat, drought, salt, and light), they serve as hubs in cellular reactions to signal and respond *via* retrograde signaling (Bobik & Burch-Smith, 2015). Once there are mutated genes in the cp genome, they might play a pivotal role in the plant's adaption to a varied environment.

In this study, we employed genome sequencing technology and bioinformatics to conduct a comparative analysis of whole cp genomes between *K. candel* samples collected from Malaysia, Thailand, and Bangladesh, and *K. obovata* samples collected from the coasts of southeastern China. The results obtained from the present study will advance our understanding of differential adaptation to coastal environments across the world conferred by variations in the cp genomes of closely related mangrove species. On a finer scale, this study will also provide a foundation for further unveiling the differential adaptation mechanisms to light and temperature in *K. obovata* and *K. candel*.

2 Materials and methods

2.1 Sampling sites and sample collection

Plant materials of *Kandelia* species were collected from China, Bangladesh, Thailand, and Malaysia, with 18 samples of *K. obovata* and seven samples of *K. candel*. Samples of *K. obovata* were collected from Zhejiang (28°34'N,121°19'E), Fujian (27°29'N, 120°29'E to 23°55'N,117°24'E), Guangdong (19°51'N,110°37'E to 21°37'N,109°47'E), and Hainan (18° 45'N,109°10'E to 19°51'N,109°15'E) Provinces in China. Among these, two planted sites, i.e., Yueqing (YQ), Zhejiang Province, and Ledong (LD), Hainan Province, were included. Samples of *K. candel* were collected from Khulna, Bangladesh (22°16'N,89°26'E); Ranong, Thailand (10°10'N,98°43'E); and Pahang, Malaysia (3°81'N,103°34'E), which are all located in the west of the South China Sea. Sampling details can be found in Figure 1 and Supplementary Table S1. One to three individual(s) per population were sampled where the population number was 11. The Sampled individuals were at least 10~15 m apart. Healthy young leaves were collected, dried in silica gel and stored at −20℃ before use. A total of 25 individual samples were used for analysis in this study.

2.2 DNA extraction and sequencing

The whole genomic DNA of 25 individual samples was individually extracted from leaf tissues using the cetyltrimethylammonium bromide (CTAB) method (Doyle, 1991). The extracted DNA was dissolved in 60 μL of TE buffer. High-quality DNA (concentration >35 ng/μL) was used in Illumina resequencing. *Kandelia* genome sequencing was performed using the Illumina Hiseq X Ten platform with 150 bp paired-end reads.

2.3 Chloroplast genome assembly, gene annotation, and codon usage

Before the *de novo* assembly of the cp genome, quality control of the raw paired-end reads was tested using Trimmomatic v0.40 (Bolger et al., 2014). High quality reads were mapped to cp seed based on bowtie2 v2.3.2 (Langmead & Salzberg, 2012) software with default parameters. The qualitatively assessed paired-end reads from bowtie2 were assembled to produce the cp genome with NOVOPlasty v4.3.1 software (Dierckxsens et al., 2020) using the published cp genome of *Kandelia obovata* (GenBank Acc. No. MN117072) as a reference (Yang et al., 2019). GetOrganelle software (Jin et al., 2020) was also applied to complete and double-check the assembly of the cp genomes. The coverage of reads in the nuclear genome was measured, which used passed Trimmomatic reads mapping to published nuclear genome based on bowtie2 v2.3.2 (Langmead & Salzberg, 2012) and the mapped genome reads were counted to coverage depth. We also used the same methods to employed the coverage depth of *Kandelia* cp genoma. The complete cp genome was annotated and corrected using Geseq (Tillich et al., 2017) and CPGAVAS2 (Shi et al., 2019). The circular gene map was visualized using OGDRAW (Greiner et al., 2019). In order to ensure the reliability and accuracy of assemble and annotation results, we have carried out manual

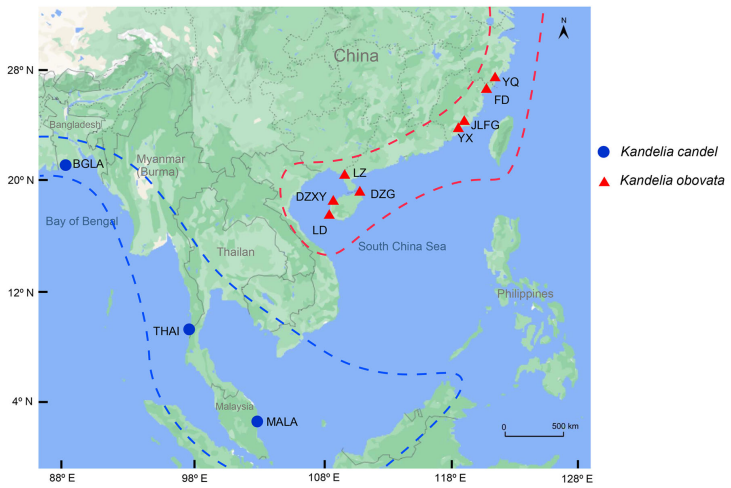


FIGURE 1
Geographic distribution of 25 *Kandelia* samples collected from 4 different countries including 11 sampling sites for sequencing and assembly the whole chloroplast genomes. The blue solid circle represented *Kandelia candel* species and the red tangle represented *Kandelia obovata* species.

comparing our results with published cp genomes of *K. obovata* (Yang et al., 2019a and 2019b). Two genes, *psbI* and *psbK*, were absent in the previously published cp genome of *K. obovata* (GenBank Acc. No NC042718, MN313722), but they were located in the cp genomes of the mangrove Rhizophoraceae family, such as in *Rhizophora apiculata* (GenBank Acc. No. MT129631.1) and non-Rhizophoraceae families, such as *Avicennia marina* in the Acanthaceae family as shown in our previous research (GenBank accession number: MT108381), as well as in our present study. Therefore, we designed primers around gene *psbK* and performed PCR amplification of cp genomes for further confirmation. The sequence of ATATTTGA ATTTGAATTGAGTTTCGGT was used as *psbK*-around-F primers. The sequence of GGTTTGTTGGATGTGCTGTGA was used as *psbK*-around-R primers. The annotated chloroplast genome sequences for the 25 samples of *Kandelia* have been deposited in the GenBank database under accession numbers successively from ON969308 to ON969332 (Supplementary Table S1). To identify codon usage patterns, all coding sequences (CDS) were subsequently used for the estimation of relative synonymous codon usage (RSCU) through the CUSP program with EMBOSS v6.5.7 with default parameters (Rice et al., 2000).

2.4 SNP calling, PCA, and phylogenetic tree

Chloroplast genome sequences of the 25 individuals from both *K. candel* and *K. obovata* were comparatively analyzed using BWA v0.7.12 with default parameters (Li & Durbin, 2009), which were aligned to the longest cp genome of *K. obovata* (YQ-2). Single-sample variant calling was performed with the Genome Analysis Toolkit (gatk v4.2.2.0) (McKenna et al., 2010) with default parameters. Gatk SortSam, gatk MarkDuplicates, gatk HaplotypeCaller and gatk CombineGVCFs were combined for SNPs calling. High-quality SNPs were kept and filtered using vcftools software (Danecek et al., 2011) with the following parameters: max missing of 0.6 and minQ of 30. The EIGENSOFT v7.2.1 package (https://github.com/gurinovich/PopCluster) was used to perform PCA, and EIGENSTRAT (Alexander et al., 2009) was performed on linkage disequilibrium (LD)-pruned pseudomolecule SNPs. The p-distance matrix was calculated using VCF2Dis (v1.47) (https://github.com/BGI-shenzhen/VCF2Dis) with the filtered SNP set from the 25 *Kandelia* accessions. A neighbor-joining tree was reconstructed using the UPGMA method, and MEGA (v5.2) (Kumar et al., 2018) was used to visualize the tree. To determine the phylogenetic relationship within the genus *Kandelia*, complete cp genomes were compared with the 25 samples among 76 shared single copy protein-coding genes to build a phylogenetic tree. The 76 genes were explored by ORTHOMCL v6.11 (Chen et al., 2006). The amino acid sequence alignments were based on MAFFT v7.487 (Katoh & Standley, 2013) with default parameters. The phylogenetic trees of *K. candel* and *K. obovata* were established with the UPGMA method by utilizing 76 shared single-copy genes within amino acid sequence (Figure 2A) based on MEGA (v5.2) (Kumar et al., 2018). To understand the phylogenetic position of the two well-differentiated geographical sets of *Kandelia* in mangroves, 13 published cp genomes of seven mangrove genera (*Rhizophora, Bruguiera, Lumnitzera, Laguncularia, Sonnerati, Avicennia*, and *Scyphiphora*) were downloaded from NCBI. *Arabidopsis thaliana* (NC_000932.1) (Sato et al., 1999) was used as the outgroup. ORTHOMCL v6.11 (Chen et al., 2006) was applied to identify orthologous gene families in 40 cp genomes. Single-copy orthologues were identified, and the BLASTP E-value was below $10e^{-5}$. Using 38 cp single-copy protein-coding genes (*accD*, *atpB*, *atpE*, *atpF*, *atpH*, *atpI*, *ccsA*, *cemA*, *matK*, *ndhA*, *ndhC*, *ndhD*, *ndhE*, *ndhG*, *ndhH*, *ndhI*, *ndhJ*, *ndhK*, *petB*, *petD*, *petG*, *petL*, *petN*, *psaA*, *psaB*, *psaC*, *psaJ*, *psbB*, *psbC*, *psbD*, *psbE*, *psbH*, *psbJ*, *psbL*, *psbM*, *psbN*, *psbT* and *rbcL*), a phylogenetic tree was reconstructed by amino acid sequence. Multiple sequence alignments of shared gene datasets were generated with MAFFT v7.487 (Katoh & Standley, 2013) with default parameters. The UPGMA method was used to infer the phylogenetic tree using MEGA (v5.2) (Kumar et al., 2018), and the parameters were adjusted for the tree with 1000 bootstrap replicates using the JTT matrix-based method with the units of the number of amino acid substitutions per site.

2.5 Comparative chloroplast genome structure analyses

Tandem repetitive sequences were determined using the online program Tandem Repeats Finder (v 4.09) (Benson, 1999) according to the following criteria: match value of 2, mismatch value of 7, delta value of 7, match probability of 85, indel probability of 10, minscore value of 50, and maxperiod value of 500. The tuple sizes were 0, 4, 5, and 7, and the tuple distances were set to 0, 29, 159, and 500. The program REPuter (Kurtz et al., 2001) was used to identify and determine the locations and sizes of forward, reverse, palindrome, and complement sequences having the following parameters: minimum of 30 bp, sequence identity greater than 90%, and maximum computed repeats of 4500. A perl program MicroSAtellite (MISA v2.1) (Beier et al., 2017) identification tool was used to detect simple sequence repeats (SSRs) in the 25 cp genomes. In this study, only perfect repeats were selected for analysis with the following parameters: basic motifs (1~6 bp) and a minimum repeat length of 8 bp (for mono- and di-), 12 bp (for tri- and tetra-), 15 bp (for penta-), 18 bp (for hexa-), and the minimum distance between two SSRs was set to 100. The variations between *K. candel* and *K. obovata* were identified by Genome Varscan (parameters: Diff OneInHundred, VarRange

100) and then the primers were designed by Batch Target Region Primer Design with default parameters, the above two tools were used under TBtools v1.099 (Chen et al., 2018) software.

To determine the sequence divergence of the *Kandelia* cp genomes among the 25 samples, the online genome comparison tool mVISTA (https://genome.lbl.gov/vista/index.shtml) was used with the *K. obovata* (YQ-2) annotation as the reference. The default parameters were set to align the cp genome in Shufe-LAGAN mode, and the sequence conservation profile was visualized using an mVISTA plot. Based on the comparative genome size results, *K. candel* (MALA-2) and *K. obovata* (YQ-2) were selected to visualize the cp genome gene order and the collinear blocks between the two species under *Kandelia*. The comparison was performed using Mauve v2.3.1 (Darling et al., 2004) with default iterative alignment, seed weight, sum of pairs LCB scoring, and LCB settings. DnaSP v5.10 (Rozas et al., 2017) was applied to determine the level of nucleotide diversity (*Pi*) among 25 samples, with the *K. obovata* (YQ-2) cp genome as the reference. When DnaSP6 calculated the *Pi* value, the step size was set to 650 bp, and the slide window size was 650 bp. The intraspecific *Pi* values of *K. candel* and *K. obovata* were also calculated using these methods. We extended the region with high *pi* value (*Pi* >0.01) up and down 1000bp to find the gene closest to this region for further analysis. The ratio of the number of non-synonymous (Ka) substitutions to the number of synonymous (Ks) substitutions (Ka/Ks) of each protein-coding gene was estimated using perl script ParaAT2.0 (Zhang et al., 2012b) with muscle v3.8.31 (Edgar, 2004) and KaKs_Calculator2.0 (parameters: -c 11 -m MS) (Wang et al., 2010).

2.6 Structural modeling and molecular dynamics simulation of *ndhD* and *atpA*

Each pair of genes of *Kandelia* was aligned using CLC Main Workbench 6 software (https://digitalinsights.qiagen.com/). Loci with variations in the species-specific genes with non-synonymous mutations were the focus of our attention between *K. candel* and *K. obovata*. We searched the genes with mutations and identified mutated sites located in the domain area with SMART (Letunic et al., 2021). The *ndhD* and *atpA* genes met the criteria. The structural models of *ndhD* and *atpA* were then built.

Templates for homologous modeling were searched *via* BLASTP (Johnson et al., 2008) on the Protein Data Bank (PDB) database (Burley et al., 2021) with a criterion of sequence identity > 30% (more is better). With a comprehensive consideration of E-value and query coverage, an optimal template was selected by PDB ID 1FX0_A (Groth & Pohl, 2001) for *atpA* and 6HUM_D for *ndhD* (Schuller et al., 2019). Model building for *K. candel* proteins was then carried out on MOE software (https://www.chemcomp.com/Products.htm) with default settings. *K. obovata* proteins were obtained by modifying the *K. candel* model with corresponding mutation positions. MOE Protein Builder was used for specific amino acid mutation followed by an energy minimization choice of "Selected Sidechain + Tether BB".

Molecular dynamics (MD) simulations were performed using the GROMACS v2021.5 (Van Der Spoel et al., 2005) with GROMOS 54A7 force field (Nathan et al., 2011). Systems were solvated with a water model (spc216.gro file) in a dodecahedron box with a boundary distance of 1.0 nm to the box edge. To maintain a neutral condition, counter-ions were added to the systems (6 Na^+ for *atpA*, 2 Cl^- for *ndhD*). The initial structures were then optimized by an energy minimization process using the Steepest Descent method until the energy gradient was ≤ 10 kJ mol^{-1} nm^{-1}. Subsequently, a two-stage equilibration process was conducted. The first 100-ps NVT equilibration for temperature control and the second 100-ps NPT equilibration for pressure control. The final MD simulations lasted 50 ns at a constant temperature of 300 K and a constant pressure of 1 atm for each system. Root Mean Square Deviation (RMSD), root-mean-square fluctuations (RMSF), and b-factor value of the residues were calculated from the MD trajectory files using the embedded commands in GROMACS. Pymol v2.4.1 (https://github.com/schrodinger/pymol-open-source) was used for protein structure visualization and mapping the b-factor value to the corresponding residue on the graph. As the b-factor value went from small to large, the colors changed in the following order: grey - blue -yellow - orange - red.

To study the interaction between ATP-A and ADP, molecular docking was carried out on MOE software with a General docking scenario. The potential binding site was selected based on the result of the Site Finder application in MOE.

To explore the consistency of ATP-A and NDH-D sequences changes and genes differentiation in mangrove plants and model angiosperms *Arabidopsis thaliana*. The proteins sequence of APA-A and NDH-D of six different species were compared and visualized with CLC Main Workbench 6 software (https://digitalinsights.qiagen.com/). Based on the comparison results, the phylogenetic trees of these two proteins were established respectively, the parameters as follows: Algorithm NJ, replicates 1000.

3 Results

3.1 General features of *Kandelia* complete chloroplast genomes

A total of 25 samples of *Kandelia* were used to obtain 5~10 Gb raw reads with a mean coverage of ~25× to 50× of whole genomes and 2298× to 3484× of cp genome base

coverage. The amplified sequences were consistent with the assembly results, which illustrated that the *psbI* and *psbK* genes were present in *Kandelia* cp genomes (Supplementary Figure S1), and the assembly of the cp genomes was correct (Figure 3). The circular maps of the 25 cp genomes are shown in Figure 3. The 25 cp genomes ranged from 165,247 to 168,262 bp in length. The total sizes of *K. candel* ranged from 165,247 to 166,729 bp, while the sizes of *K. obovata* ranged from 168,070 to 168,262 bp. The length of *K. obovata* was approximately 2 kb longer than that of *K. candel* on average. All cp genomes had a circular assembly with a typical quadripartite structure, which was composed of large and small single-copy (LSC and SSC) regions and two inverted repeats (IRs) (Figure 3, Table 1). The LSC length showed the greatest difference between the two species, being 92,380~93,683 bp for *K. candel* and 94,711~94,908 bp for *K. obovata*, which resulted in the differential length of the two species. GC contents in both LSC (38.2~8.3%) and SSC (28.1~28.6%) were slightly lower in *K. obovata* (Supplementary Table S4).

The LSC region comprised 59 protein-coding genes and 22 tRNA genes, and the SSC region comprised 11 protein-coding genes and one tRNA gene in the 25 *Kandelia* cp genomes. Eighteen genes contained introns, of which 15 genes (*trnK*-UUU, *rpoC1*, *atpF*, *trnG*-UCC, *trnL*-UAA, *trnV*-UAC, *petB*, *petD*, *rpl2*, *ndhB*, *trnI*-GAU, *ndhA*, *trnQ*-UUG, *trnS*-GGA and *trnA*-UGC) contained one intron, while two genes (*ycf3* and *clpP*) possessed two introns (Supplementary Table S2). The *trnk*-UUU gene featured the longest intron (2595~2607 bp), which contained the *matK* gene. The shortest intron was located in *trnL*-UAA (563~569 bp). The *rps12* gene was a trans-spliced gene with the 5'-end situated in the LSC region, with one exon, and the 3'-end situated in the IR region.

3.2 Codon usage analysis

The ATT codon was the most abundant in the *Kandelia* cp genomes (42.87%). The TGA codon was the least used in *Kandelia* cp genomes (2.69%). The analysis of relative synonymous codon usage (RSCU) values revealed the dominance of 20 amino acids. RSCU values were computed for *K. candel* and *K. obovata* cp genomes based on their protein-coding sequences. Supplementary Figure S2 shows the codon content of 20 amino acids and stop codons in all protein-coding genes in the cp genomes of the two species. The coding regions of *K. candel* were composed of 24,542, 24,680, and 24,690 codons. In *K. obovata*, the coding regions were composed of 24,674, 24,681, and 24,680 codons. The most prevalent amino acid was leucine, with 2,579~2,592 codons in *K. candel* and 2591 codons in *K. obovata*, while the rarest was cysteine, with 306 codons in *K. candel* and 307 codons in *K. obovata*, showing the sequence diversity between the two species. When codons with no preference value were set to 1.00, codons for leucine, serine, and arginine were the most abundant (RSCU = 6), while those for tryptophan and methionine were the least abundant (RSCU = 1) (Supplementary Figure S2). In addition, nearly all A/T-ending codons had RSCU values >1, whereas G/C-ending codons had RSCU values <1.

3.3 SNPs analysis and construction phylogenetic trees

There were 1522 indels in the 25 cp genomes of *Kandelia*, comprising 112 (7.36%) SSR-related indels and 1410 (92.64%) non-SSR-related indels. In total, 73.52% of the indels were present in 1141 intergenic space regions, while 2.46% of the indels were located in exons and 24.02% were present in the introns (Figure 4A). The *atpE-atpB-rbcL* regions contained 60

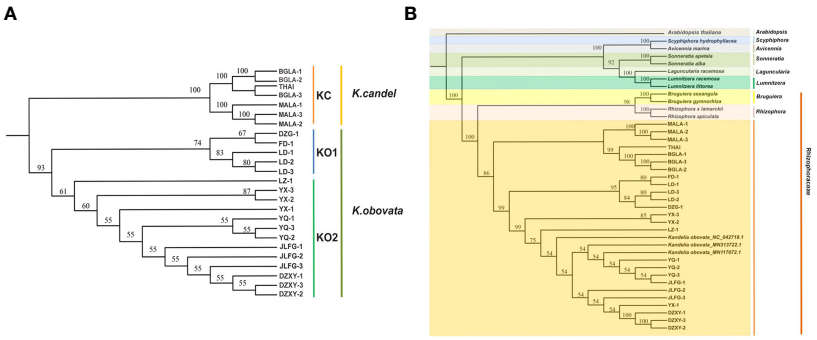


FIGURE 2
Phylogenetic tree. **(A)** The *Kandelia* tree constructed using the UPGMA model based on the 76 single-copy genes in the 25 whole cp genomes. **(B)** The tree constructed using the UPGMA model based on 38 homologous pair genes in 13 mangrove species including 40 samples.

10.3389/fpls.2022.1075353

TABLE 1 The basic chloroplast genome information of the 25 *Kandelia* accessions.

Characteristics	*K. candel*	*K. obovata*
Total Number of Raw reads	7,183,966,500 - 8,404,770,000	5,793,072,000 - 10,367,439,000
Total Number of Mapped read	2,899,354 - 3,771,743	2,605,052 - 3,949,533
Percent of chloroplast genome reads (%)	0.04% - 0.045%	0.038% - 0.045%
Chloroplast genome coverage (X)	2,558 - 3,328	2,298 - 3,484
Total size (bp)	165,247- 166,729	168,070 - 168,262
LSC length (bp)	92,380 - 93,683	94,711 - 94,908
IR length (bp)	266,25 - 266,87	266,47 - 266,72
SSC length (bp)	196,70 - 199,25	199,64 - 200,39
Total genes	128	128
Protein coding genes	83	83
tRNA genes	37	37
rRNA genes	8	8
Overall GC content (%)	36.10 - 36.30	36.1
GC content in LSC (%)	38.3	38.2
GC content in IR (%)	42	42
GC content in SSC (%)	28.10 - 28.60	28.10 - 28.20
Accession number	ON969322 - ON969325, ON969308,ON96939,ON969332	ON969310 - ON969321, ON969326 - ON969331

indels, followed by *trnS* (GCU)-*rps4-trnT* (UGU) with 44 indels and *trnD* (GUC)-*psbM-petN* with 33 indels. The single nucleotide site generally displayed a high frequency of SSR-related indels in the present study; however, we also found five indels located in the *trnS* (GCU)-*rps4-trnT* (UGU) region, which were 13~56 bp in length. All SSR-related indels were A/T-type SSRs. Three SSR-related indels were located in the coding regions, and the other 41 SSR-related indels were found in the non-coding regions. The sizes of the non-SSR-related indels ranged from 1 to 169 bp, with one bp indel (1,171) being the most common (Figure 4B). The largest indel (88 bp) in the spacer of *trnK* (UUU)–*trnQ* (UUG) was a deletion in *K. candel*. The second largest indel was in the spacer of *ndhG–ndhI* (81 bp), which was also a deletion in *K. candel*.

The relationships among the 25 samples were analyzed using principal component analysis (PCA) based on SNPs data, and a phylogenetic tree of the SNP markers in the whole cp genome was identified (Figures 4C, D). Using the longest cp genome of YQ-2 as the reference genome, the other 24 cp genomes were clustered into two main groups, KC and KO, the latter included KO1 and KO2 subgroups. The KC group belonged to *K. candel* and included all samples (THAI, BGLA-1, BGLA-2, BGLA-3, MALA-1, MALA-2, and MALA-3) of *K. candel*. The KO1 and KO2 groups belonged to *K. obovata*, which consisted of DZXY1-3, JLFG1-3, YX1-3, YQ-1, YQ-3, and LZ-1 in the KO1 group and LD-1, LD-2, LD-3, FD-1, and DZG-1 in the KO2 group. The first two principal components comprised 68.32% of the total variance, with PC1 reflecting the variability of the *K. candel* and *K. obovata* groups.

The phylogenetic trees of *K. candel* and *K. obovata* within the genus *Kandelia* were established with the UPGMA method by utilizing 76 shared single-copy protein genes (Figure 2A). The phylogenetic trees showed that the individuals of *K. candel* clustered in one clade as a KC group. *Kandelia obovata* individuals were gathered in another clade, and samples were gathered further into the KO1 and KO2 subgroups, which was consistent with the phylogenetic tree constructed based on SNPs (Figure 4D). The KC branch of *K. candel* was subdivided into two clades: MALA1-3 for one clade and BGLA1-3 together with THAI for another clade. The KO1 and KO2 groups of *K. obovata* branches showed that most samples from the same location were more closely related; however, some samples collected from neighboring provinces or even far from each other also showed closer kinship. For example, LD 1-3 and DZG-1, both from Hainan Province, were clustered with FD-1 from Fujian Province, and YQ 1-3 from Zhejiang Province was grouped with JLFG-1 from Fujian Province. This might be because Fujian Province was one of the major collection areas for propagules of *K. obovata* in artificial introductions or transplants.

10.3389/fpls.2022.1075353

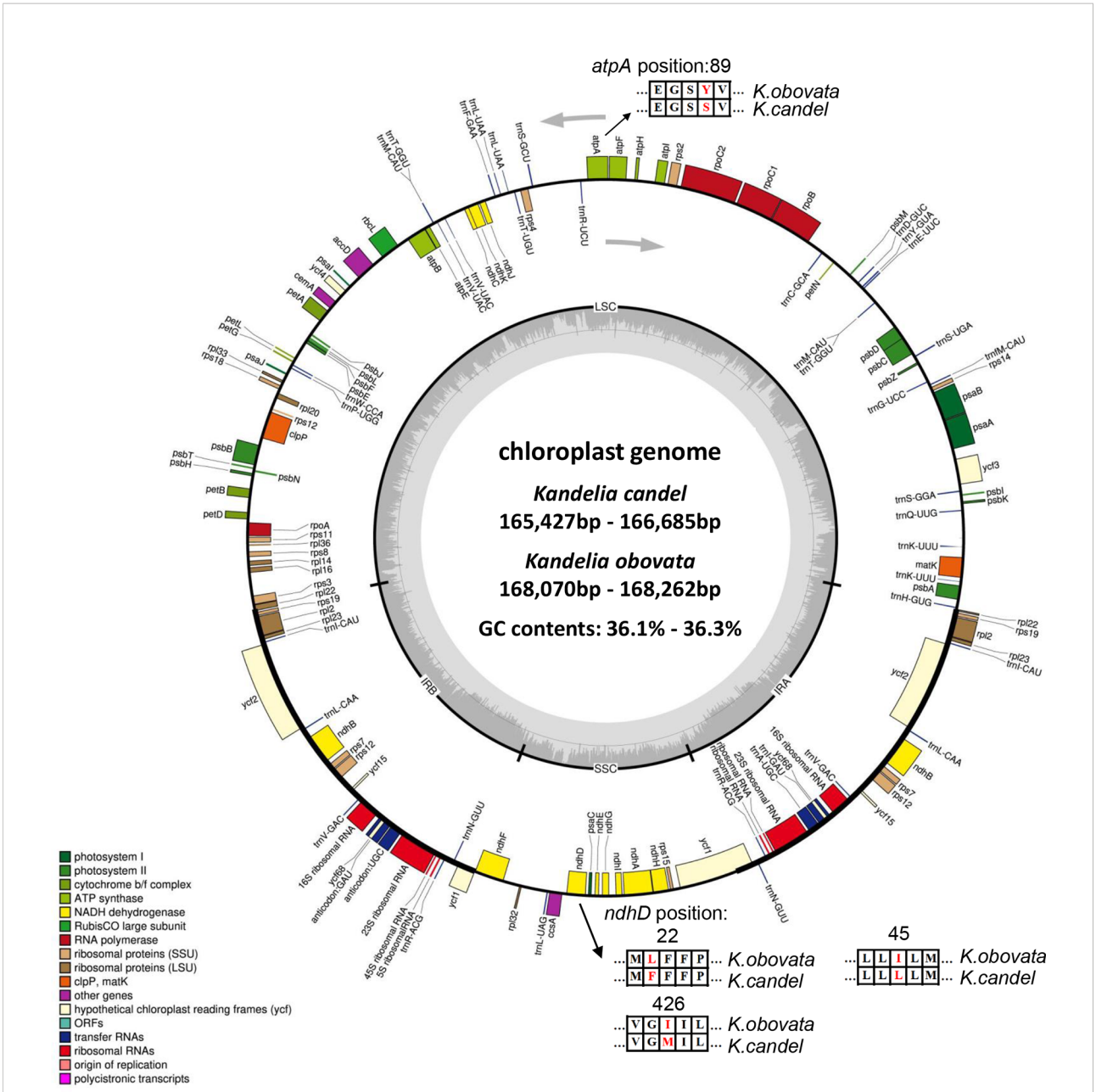


FIGURE 3
Gene maps of the chloroplast genomes of *Kandelia*. Genes shown outside the outer circle are transcribed clockwise, and genes shown inside the circle are transcribed counterclockwise. Genes belonging to different functional groups are color coded. The dashed area in the inner circle indicates the GC content of the chloroplast, and the light gray area corresponds to AT content of the chloroplast. Point mutations of *atpA* and *ndpD* gene were shown between *K. obovata* and *K. candel*.

In addition, a phylogenetic tree based on 38 homologous pair genes was constructed among 7 genera, 13 mangrove species, including 39 samples (Figure 2B). The results illustrated that *Kandelia* was closely related to *Bruguiera* and *Rhizophora* under the Rhizophoraceae family, but they were well differentiated into two species. In terms of the 25 *Kandelia* samples, this result also coincided with the phylogenetic trees constructed with SNPs (Figure 4D) and 76 shared single-copy genes in these cp genomes (Figure 2B), all illustrating that the genus *Kandelia* consisted of two distinct species.

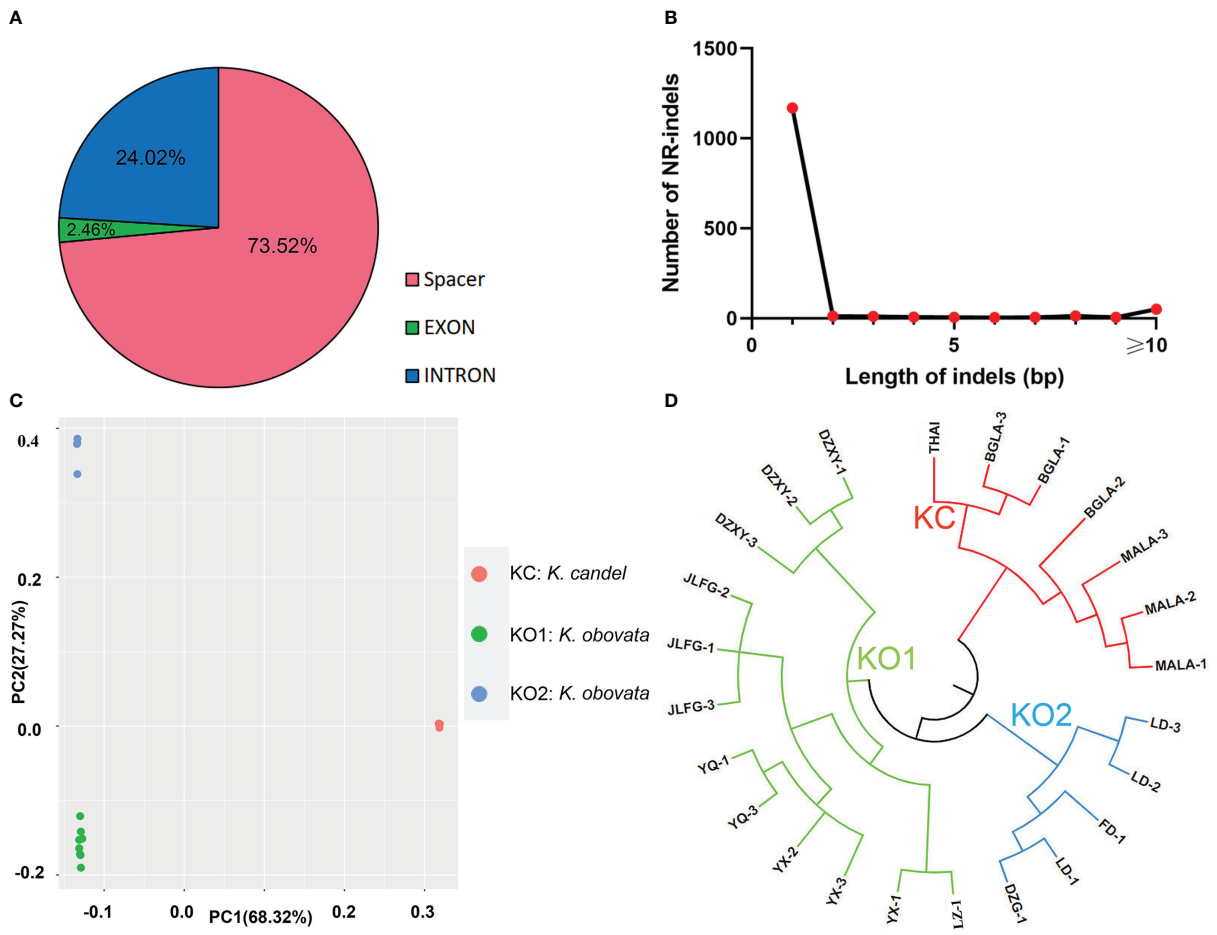


FIGURE 4
Analyses of indels in the chloroplast genomes. **(A)** Frequency of different indels types and locations. **(B)** Number and size of non-SSR-related indels in the *K. obovata* genomes. **(C)** The PCA result of SNPs. **(D)** The phylogenetic tree based on SNPs data.

3.4 Comparative analysis of cps genomes structural variations

The distribution of long repeats in *Kandelia* cp genomic sequences was analyzed and summarized, as shown in Supplementary Figure S3A. There were three types of repeats: tandem repeats (TRs), forward repeats (FRs), and palindromic repeats (PRs) (Supplementary Figure S3A). By searching the repeats of each of the 25 cp genomes, the number of repeats was significantly different in TRs and PRs between *K. candel* and *K. obovata*. There were 2~10 TRs in *K. candel* and 17~22 TRs in *K. obovata* (Supplementary Figure S3B). The average number of TRs in *K. candel* was 11, which was significantly lower than that in *K. obovata* (7). In terms of FRs, the mean number was 37 in the *K. candel* cp genomes, whereas it was 59 in *K. obovata* genomes. In contrast, the PR numbers in *K. candel* were greater than those in *K. obovata*, at 5~8 and 2~4, respectively. The highest number of repeats (88) was found in the LD-3 cp genome of *K. obovata*, with 19 TRs, 66 FRs, and three PRs. Among these repeats, 47 of the forward repeats were found with 28~46 bp, six repeats with 47~64 bp, five repeats with 65~83 bp, four repeats with 84~102 bp, three repeats with 103~120 bp, and one repeat with length longer than 121 bp (Supplementary Figure S3C). The lowest number of repeats (a total of 41) was found in the BGLA-1 cp genome of *K. candel* (nine TRs, 27 FRs, and five PRs), of which 26 of the FRs were found with 28~46 bp, 14 repeats had length lower than the average value (40) of *K. obovata*, and one repeat had a length of 47~64 bp (Supplementary Figure S3A, C, D).

The distribution of three types of SSRs, namely mononucleotides, dinucleotides, and trinucleotides, is shown in Supplementary Figure S4A. The total number of SSRs were 73 to 78 with an average density from 456.25 SSRs/Mb to 487.5 SSRs/Mb of *K. obovata*. In *K. candel*, the SSRs density were from 456.25 SSRs/Mb to 506.25 SSRs/Mb (73~81). AAT repeats were found in *Kandelia*, and there was only one

trinucleotide type duplication per cp genome. In addition, the SSR sequences of the whole cp genome displayed a prevalence of AT-rich mononucleotides (98~100%) and dinucleotides (100%).

The mononucleotide SSR was found to be the most abundant, followed by dinucleotides and then trinucleotides, in both species. They accounted for 91.86%, 6.82%, and 1.32% of total SSRs in *K. obovata* individuals, on average, while they accounted for 92.33%, 6.37%, and 1.30% in *K. candel* individuals (Supplementary Table S3). Most of the SSRs in the 25 cp genomes were found in the LSC region (Supplementary Figure S4B). A significant difference in SSR distribution between *K. candel* and *K. obovata* was also found in the LSC region. The number of SSRs in the LSC region was 53 (YX-1) to 61 (MALA-1, 2, and 3). The average SSR number in *K. candel* (59) was greater than that in *K. obovata* (56). The most mononucleotide SSRs distributed in the LSC region in *K. candel* were 54, on average, whereas there were 50 in *K. obovata*.

Similarly, SSRs in the SSC region were mostly mononucleotides, but SSRs as dinucleotides were also found in *K. candel* individuals. The number of SSRs in the SSC region was 9~12 (average of 14.37%) in *K. obovata*, which is higher than that in *K. candel* (i.e., 8~11, average of 11.99%). In the IRa region, four to six mononucleotide SSRs were found. The least SSRs were found in the IRb region with four mononucleotides, except MALA-2 with five mononucleotides. The DZXY cp genome of *K. obovata* had one dinucleotide SSR in the IRb region, which was not found in any of the other samples. Molecular markers to distinguish these two species were shown as in Supplementary Table S5.

Using the longest *K. obovata* (YQ-2) cp genome as the reference, the comparative sequence analyses exhibited high sequence similarities and high gene structure order consistency among the 25 cp genomes (Supplementary Figure S5). The co-linearity in gene placement between the cp genomes of *K. candel* and *K. obovata* was analyzed with MALA-2 and YQ-2 as representatives. The results indicated that gene clusters had no changes, whether in the single-copy regions or in inverted repeat regions between the two species, showing high conservation at the whole cp genome level (Figure 5). However, a gap of 1,149 bp in YQ-2 was detected in the LSC region (Figure 5). The gap region was very rich in AT (90.37%).

The nucleotide diversity (*Pi*) value was calculated using the DnaSP program to evaluate the mutation hotspots in the 25 cp genomes (Figure 6). The results illustrated that the Pi values varied from 0 to 0.04 in the peer window of the 25 *Kandelia* cp genomes. Seven of these loci, i.e., *atpA* (0.0156), *ndhD* (0.0148), *matK* (0.0152), *rps4* (0.0402), *atpB* (0.0228), and *rbcL* (0.0157), were located at high values (*Pi* >0.01) area. Furthermore, sequence coherence analysis showed that the *atpA* gene in the LSC region and the *ndhD* gene in the SSC region displayed two distinct patterns between *K. candel* and *K. obovata*, which aroused our interest in further comparison. In addition, the *matK* gene and the *rps4* gene in the LSC region displayed a consistent pattern in all samples of *K. candel*. The *atpB* gene and the *rbcL* gene in the LSC region displayed a consistent pattern in all samples of *K. obovata*. The *Pi* value of intraspecific cp genome nucleotide diversity in *K. candel* was between 0 and 0.023, while it was 0~0.015 in *K. obovata*. The Ka and Ks values estimated for each protein-coding gene showed that they were in the range of 0 to 0.5, and none of the Ka/Ks values for these genes were greater than 1 in the present study. In particular, the Ka/Ks value of *atpA* gene is 0.10 and that of *ndhD* gene is 0.26. (Supplementary Figure S6, Supplementary Table S3).

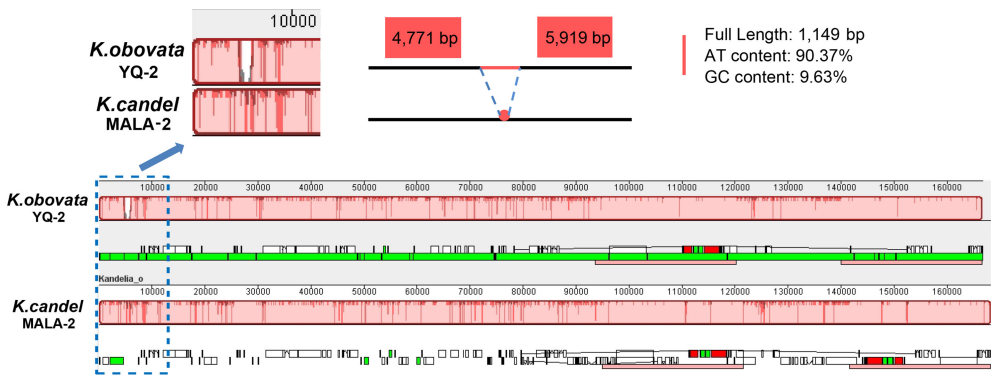


FIGURE 5
Gene map comparison between *K.candel* and *K.obovata* chloroplast genomes aligned using Mauve, showing a big 'gap' of 1,149bp with rich A and T in LSC regions in *K. obovata*.

10.3389/fpls.2022.1075353

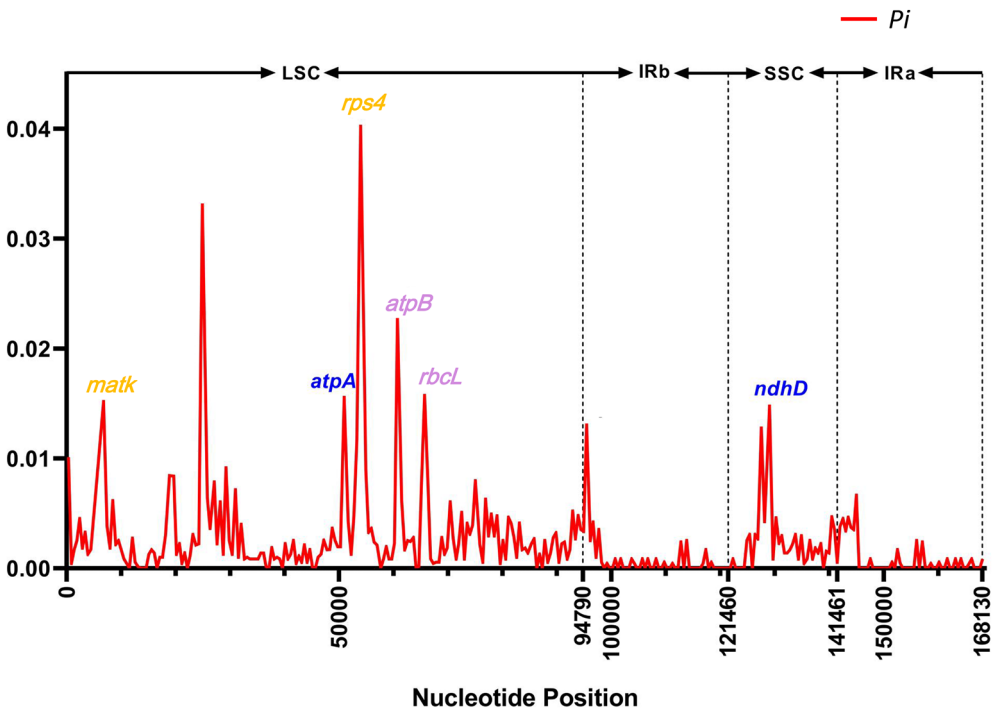


FIGURE 6
The *Pi* value in sliding-window analysis of the whole chloroplast genomes. The genes highlighted in blue color showcase the SNPs between *K. candel* and *K. obovata*. The genes highlighted in orange color were SNPs within *K. candel* species. The purple color highlighted genes were SNPs within *K. obovata* species.

3.5 Structure analysis and molecular dynamics simulation of NDH-D and ATP-A proteins

Analysis of the SNPs between *K. candel* and *K. obovata* showed that the *ndhD* and *atpA* genes had some SNPs located in highly variable regions (Figure 6). All the mutation sites of proteins NDH-D and ATP-A were identified in the domain area using the SMART program (Supplementary Figures S7A, S8A). To explore the influences of mutation sites of NDH-D and ATP-A on protein structures, we carried out homology modeling, molecular docking, and a followed-up molecular dynamic (MD) simulation. b-factors (Supplementary Figure S7B), RMSD (Supplementary Figure S8B), and RMS fluctuation (Supplementary Figure S8C) were calculated for better understanding.

The NDH complex D sub-protein (NDH-D) was reconstructed with the *K. candel* protein model by selecting the appropriate homologous protein, and *K. obovata* proteins were obtained by modifying the *K. candel* NDH-D protein model with corresponding mutation positions (Figure 7). As shown in Figure 7C, Phe22, Leu45, and Met426 of NDH complex D sub-protein (NDH-D) in *K. candel* were substituted by Leu22, Ile45, and Ile 426 in *K. obovata*. Clearly, due to the sense mutations of the protein amino acids in *K. obovata*, the protein structure changed accordingly (Figures 7A, B). After MD simulations, b-factors for residues 22, 45, and 426 were calculated. The b-factors for the 22nd, 45th, and 426th sites were all higher in *K. candel* than in *K. obovata* (Supplementary Figure S7B), indicating a more flexible conformation at these three sites in *K. candel*. Although the overall trend of the conformational changes of NDH-D in both *K. candel* and *K. obovata* was consistent (Figure 7D), the marked differences in protein conformation fluctuated around the residues 70~80, 160~180, and 450 (Figure 7E).

The structure of ATP α subunit protein (ATP-A) coded by the gene *atpA* was constructed in the same way as the NDH-D protein (Figure 8). According to the SNP result, the Tyr89 in protein ATP-A in *K. candel* was mutated to the Ser89 in *K. obovata*, (Figure 8D). The predicted structures of ATP-A were shown in Figures 8B, C. RMSF analysis after MD simulations illustrated that great differences existed in the regions from the residue 26 to 96 amino-acid in ATP-A between species (Figure 8A). As the docking results showing in Figures 8E, F, a larger absolute value of binding affinity (-6.55 kcal/mol vs. -5.56 kcal/mol) indicated a stronger interaction between ATP-A and ADP in *K. obovata*, suggesting that the α subunit protein of ATP complex may work more actively in ATP synthesis in *K. obovata*.

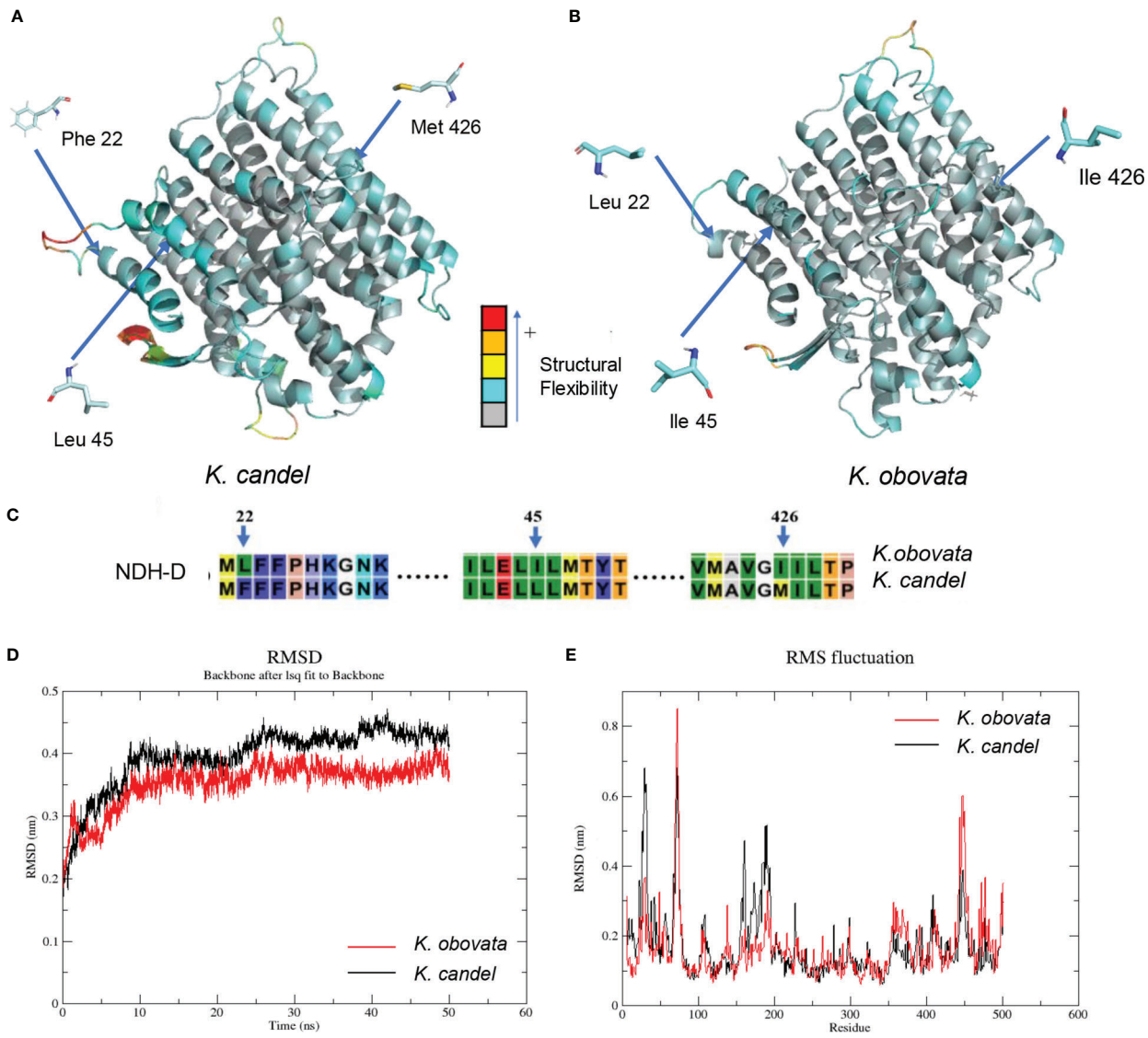


FIGURE 7
The predicated 3D structural modeling and molecular dimnamics simulation of the NDH-D protein. By setting different colors (gray to blue to yellow to orange to red) according to the b-factor values, the color in the structure is closer to red, the more flexible the structure. **(A)** The 3D model of NDH-D protein for *K. candel*. **(B)** The 3D model of NDH-D protein for *K. obovata*. **(C)** The mutation sites of protein sequence compared between *K. candel* and *K. obovata*. Molecular dynamics simulations showed with RMSD **(D)** and RMS fluctuation **(E)**.

Kandelia showed that *K. obovata* had a mutation site in ATP-A, while *K. candel* was identical at this site with others mangrove and Arabidopsis, which located in the protein domain (Supplementary Figure S9). Among these, *K. obovata* was consistent in the amino acid at the position of 22nd with *Arabidopsis* but different from *K. candel* and the other two Rhizophoraceae species, *Bruguiera sexangula* and *Rhizophora apiculata*. The sites of 22nd, 45th and 426th showed that *K. candel* and *A. thaliana* were consistent while *K.obovata* species were different from all the other compared species (Supplementary Figure S9).

The two animations for the whole 50ns MD simulations demonstrating the ATP-A-ADP binding are provided in the supplementary movie document (Supplementary Animations A, B).

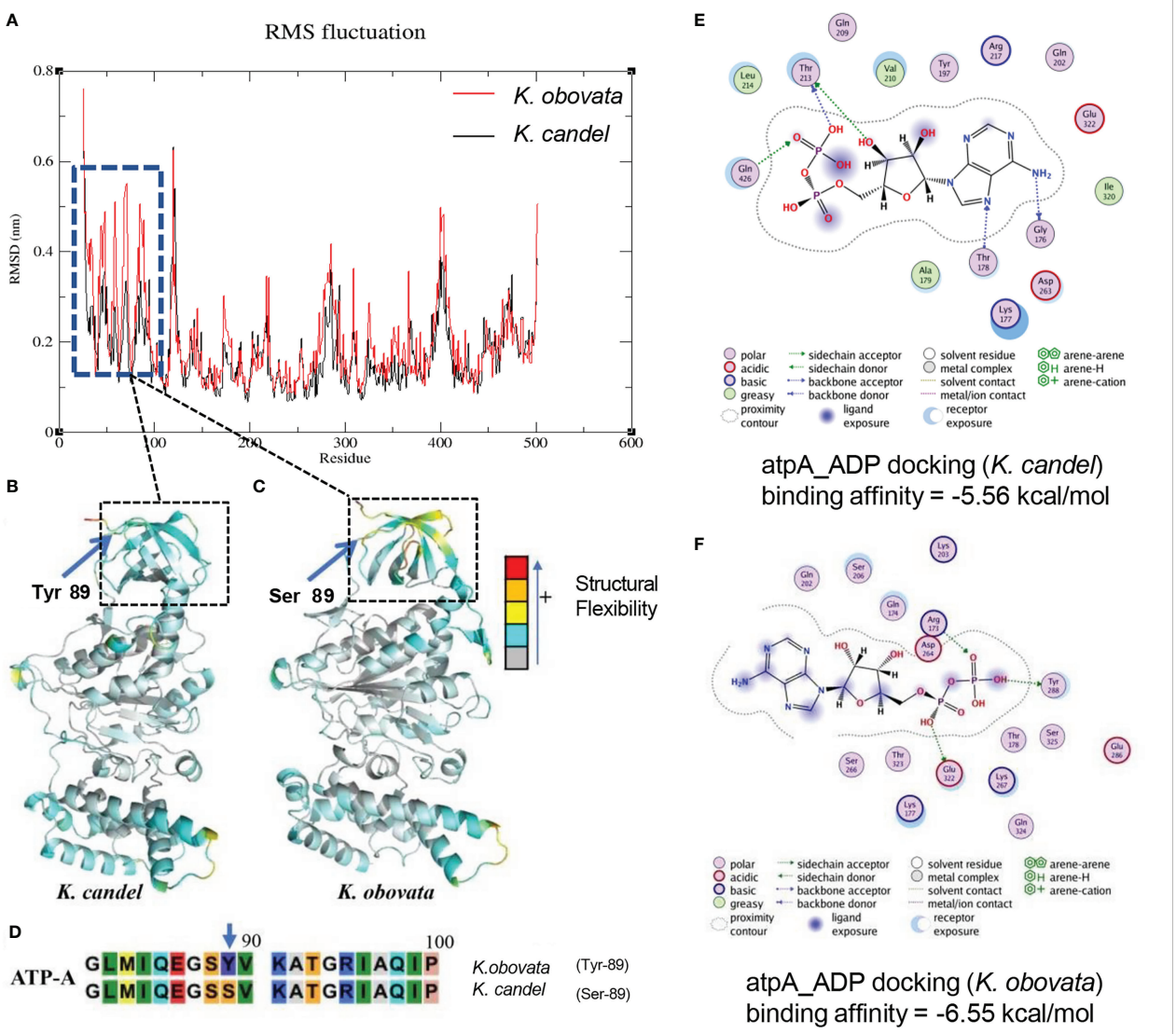


FIGURE 8
The predicated 3D structural modeling and molecular dimnamics simulation of the ATP-A protein. By setting different colors (gray to blue to yellow to orange to red) according to the b-factor values, the color in the structure is closer to red, the more flexible the structure. **(A)** The conformational variations of protein amino acid residues during the molecular dynamics simulation process showing the distinct difference in the region from 26aa to 96aa. **(B)** The 3D model of ATP-A protein for *K. candel*. **(C)** The 3D model of ATP-A protein for *K. obovata*. **(D)** The mutation at the site of the 89 aa of ATP-A protein between *K. candel* and *K. obovata*. **(E)** Molecular dimnamics simulation of ATP-A protein and ligand ADP docking exhibuted by binding affinity for *K. candel* **(E)** and *K. obovata* **(F)**.

4 Discussion

4.1 Comparative analysis of the 25 cp genomes gave support to *Kandelia* being differentiated into two distinct species

Kandelia (Rhizophoraceae) was regarded as a monotypic genus consisting of a single *K. candel* (Hou, 1958; Juncosa & Tomlinson, 1988). Based on studies on leaf anatomy, cold-resistance adaptation, chromosome number, and molecular markers from both cp and mitochondria, *K. obovata* was identified as a new mangrove species differing from *K. candel* 20 years ago (Sheue et al., 2003), but the genomic information between them remained unknown, which limited our understanding about how their genes evolved and differentiated during the process of plant speciation. This also causes confusion in scientific exchanges, as some scholars still use *K. candel* for the plant materials of *Kandelia* collected either from China or Japan (Geng et al., 2008; Enoki et al., 2009; Zhang et al., 2012a; Feng et al., 2022).

The plastid genome provides valuable information for species identification, population genetics, species differentiation, and phylogenetic analyses (Daniell et al., 2016; Kong et al., 2021; Luo et al., 2021). In this study, we obtained the high-quality cp genomes of *Kandelia* (Table 1, Figure 3), with the cp genome size ranging from 165,247 to 168,262 bp, which were almost 3 kb larger than previously published *K. obovata* cp genomes (Chen et al., 2019; Yang et al., 2019). The number of genes and arrangements were also different in comparison with previous studies. Genes *psbK* and *psbI* were assembled and annotated in our study, which were located in the aforementioned 3 kb regions, but they were absent in previous studies (Chen et al., 2019; Yang et al., 2019). To confirm the correctness of our cp genome assembly, we designed primers around gene *psbK* and double-checked the locations of *psbK* and *psbI* genes by randomly selecting seven samples with PCR amplification. The results verified that they were in the right sequence position in the cp genomes (Supplementary Figure S1). In addition, the fragment from 8,966 to 52,067 bp was in inverse order in the LSC region, which was different from the reported cp genome of *K. obovata* (Chen et al., 2019; Yang et al., 2019). This inversion has been previously reported in other cp genomes (Wei et al., 2020; Cui et al., 2021; Kim & Cheon, 2021; Wang et al., 2021).

The genomes of all samples shared 83 protein-coding genes, 37 transfer RNA genes, and eight ribosomal RNA genes. The size of the cp genome in *K. obovata*, however, was roughly 2 kb longer than that in *K. candel*, with the greatest difference in the LSC regions. The GC content in the IRs remained the same, but both LSC and SSC were slightly lower in *K. obovata* in comparison with *K. candel* (Figure 3, Supplementary Table S4).

Indels are an important class of genetic variations and play important roles in species evolution (Biju et al., 2019; Yang et al., 2019). We identified 1,522 indels in 25 cp genomes of *Kandelia*, comprising 112 SSR-related indels and 1,410 non-SSR-related indels (Figure 4). The phylogenetic tree based on the indel data of all cp genomes divided 25 samples into two separate branches. PCA analysis was in accordance with the phylogenetic tree established by the SNPs. Phylogenetic relationship analyses based on 76 homologous protein genes from the single-copy genes in these 25 cp genomes (Figure 2A) and 38 homologous pair genes among 13 mangrove species, including 40 samples (Figure 2B), provided strong support to the genus *Kandelia* consisting of two distinct species. These two phylogenetic analyses also coincided with the results obtained by the SNPs of the whole cp genome tree. Our integrative phylogenetic trees provided detailed molecular information revealing that *K. candel* and *K. obovata* are two species geographically separated by the South China Sea, as previously suggested by Sheue et al. (2003). In terms of *K. obovata* clustering into two subgroups, among the many possible causes, introduction or transplant was likely the main reason for gene homogenization. Among these, YQ in Zhejiang Province and LD in Hainan Province were the two artificial introduction places for *K. obovata*, and Fujian was the main provenance of the propagules of *K. obovata* (Li et al., 2001; Wang & Wang, 2007).

Repeat sequences are an important part of the genome and play an important role in the evolution of organisms, genetics, and regulation of gene expression. Comparison of the tandem, forward, and palindromic repeats markedly differed between *K. candel* and *K. obovata* (Supplementary Figure S3). Tandem and forward repeats in *K. obovata* were more frequent than those in *K. candel* (Supplementary Figure S3B, C). In contrast, palindromic repeats were less frequent in *K. obovata* than in *K. candel* (Supplementary Figure S3D). This also resulted from the differences in cp genome sizes between *K. candel* and *K. obovata*, due to these different repetitive sequences mainly located in the intergenic regions. Microsatellites, also known as simple sequence repeats (SSRs), are short repeating DNA sequences of one to six base pairs that are ubiquitous in the genome (Powell et al., 1996; Fang et al., 2020). Owing to the number of repeat units that may vary between individual genotypes, SSR is extremely versatile and useful for genetic analysis (Zalapa et al., 2012). In this study, more SSRs were found in *K. candel* than in *K. obovata*, especially the type of mononucleotide in the Malaysian samples (Supplementary Figure S4). Moreover, SSRs typically consist of a higher number of AT bases than GC bases (Wang et al., 2022), which is consistent with our observations and the high AT content in the nucleotide composition of *K. candel*. Given that SSRs were applied to create maps of genetic linkage, variety identification, and molecular markers (Gupta et al., 2022; Steele et al., 2006), detailed information on the SSRs identified in the present study can facilitate future research on selected target regions for more in-depth population studies between these two species within genus *Kandelia*.

4.2 Variations and non-synonymous mutations of the cp genomes between *K. candel* and *K. obovata*

Mauve analysis showed an indel with a length of 1,149 bp in the LSC region between *K. obovata* and *K. candel* (Figure 5), which explains the striking difference in cp genome sequence composition between *K. candel* and *K. obovata*. Near this region with a high nucleotide diversity value (within 2 kb upstream and downstream), there were non-synonymous mutations in the *atpA* and *ndhD* genes between *K. candel* and *K. obovata* (Figure 6). *MatK* and *rps4* had non-synonymous mutation sites in different samples of *K. candel*, while the *atpB* and *rbcL* genes had non-synonymous mutation sites in different samples of *K. obovata* (Figure 6). Moreover, the genes with a *Pi* value of >0.01 mentioned above were located in the LSC and SSC regions. Taking cp DNA *atpB-rbcL* spacer and *trnL-trnF* spacer as examples, both were used as universal cp DNA markers to

identify the two geographically distributed populations of *Kandelia* (Chiang et al., 2001). In the present study, detailed information obtained by comparison of the 25 cp genomes of *Kandelia* indicated that there were 58 variant loci in *atpB-rbcL* and four variant loci in *trnL-trnF* spacers between the two species. As the cp genome has a copy-dependent repair mechanism, which in turn ensures the conservation and stability of the two IR regions in the cp genome (Khakhlova & Bock, 2006; Wang et al., 2022), the variation in the IR region was much less than in the LSC and SSC regions. Sequence variations existed throughout the cp genomes between the *K. candel* and *K. obovata*, especially the non-synonymous mutation sites between the two species. Predicting the mutation sites in non-synonymous amino acid residues would doubtlessly provide a foundation for further research on this protein in adaptation to habitat.

4.3 Structural and functional simulation analyses of non-synonymous proteins between *K. candel* and *K. obovata*

Light intensity and temperature are the two key environmental factors for differentiating *K. candel* and *K. obovata* geographically distributed in the northern and southern banks of the South China Sea, respectively. Adaptations to these key factors are the most critical characteristics of their differential genes and corresponding protein evolution in the northward colonization of *Kandelia*. Comparative analyses of sequence divergence and mutation hotspots revealed that the *atpA* gene in the LSC region and the *ndhD* gene in the SSC region exhibited distinct patterns between *K. candel* and *K. obovata* (Figures 3, 5, 6, 7C, 8D).

The cp NDH complex is composed of 11 plastid genes (*ndh* A~J) together with at least 18 nuclear genes that function as thylakoid NAD(P)H dehydrogenase (Shen et al., 2021). The complex is involved in electron transfer from NAD(P)H to plastoquinone, protecting the plant cell against photooxidative stress and maintaining optimal rates of photosystem I (PSI) cyclic photophosphorylation (Braukmann et al., 2009; Peredo et al., 2013). NDH-dependent cyclic electron flow (CEF) provides extra ΔpH and ATP for the CO_2 assimilation, which is essential for balancing the changing demands for ATP/NADPH and regulating photosynthetic machinery in response to various environmental conditions (Zhu et al., 2001; Yukawa et al., 2005; Shikanai, 2016; Shikanai, 2020; Shen et al., 2021). Therefore, it functions as a value-feeding electron to poise the redox level of the intermediaries to optimize the rate of the cyclic electron transport in accordance with the changes in light intensity (Casano et al., 2000; Joet et al., 2002). More research has shown that *ndh* genes are of great importance in photosynthesis regulation in adaptation to harsh environments. Small changes in any of the *ndh* genes significantly decrease the photosynthesis rate (Endo et al., 1999; Silva et al., 2016). Adaptation to submersed environments is accompanied by the complete loss of the NDH complex in an aquatic angiosperm (Peredo et al., 2013). A series of T-to-C inactivating mutations occurred in *ndh* genes, which were further corrected back to T during evolution (Martín & Sabater, 2010), implying the importance of the *ndh* genes in the stability of the complex. Given the location of NDH-D in the transmembrane helix of the NDH complex, it plays a key role in maintaining the activity and plasticity of the complex (Li et al., 2013; Shen et al., 2021). Any variations in NDH-D would impact photosynthetic electron transport, energy generation, and light adaptation.

In the present study, the *ndh-D* gene showed three non-synonymous mutation sites, two of which mutated from base C in *K. candel* to base A in *K. obovata* and one from base G in *K. candel* to base A in *K. obovata*, which therefore caused the amino acids at three sites to vary from F (Phe), L (Leu), and M (Met) to L (Leu), I (Ile), and I (Ile) in *K. obovata*, showing intraspecific consistency but the obvious interspecific difference at the 22nd, 45th, and 426th amino acids in the protein sequence between the two species (Figure 7). Further comparative analysis of the structural modeling of the NDH-D protein between the two species illustrated that b-factors for the three sites were higher in *K. candel*, illustrating that the protein possessed higher flexibility at the three amino acid regions in comparison with *K. obovata* (Figure 7). As protein structural plasticity is closely related to its active function in light harvesting and photosynthetic electron transport (Tian et al., 2012; Sun et al., 2019), considering *K. candel* mainly distributed in Southeastern Asia with higher light intensity and temperature in comparison with *K. obovata* growing in Eastern Asia, we suggested this might relate to the long-term adaptation to differential light radiation in their respective habitats.

Temperature has a strong impact on protein evolution. It has been found that orthologous proteins of species evolved at different temperatures commonly exhibit distinctive differences in function and structural stability in adaptation to temperature (Fields et al., 2015; Cvetkovska et al., 2018). In chloroplasts, F_0F_1-ATP synthase (cp H^+-ATPase) is a protein complex responsible for ATP synthesis and energy generation. It contains two components with independent functions, a membrane-intrinsic CF_0, which is responsible for the proton flux across the membrane, and a membrane-extrinsic CF_1, which catalyzes the synthesis of ATP from ADP and phosphate using the energy of electrochemical transmembrane potential of protons (Nelson & Ben-Shem, 2004; Hahn et al., 2018). The core component of CF_1 is $\alpha 3\beta 3\gamma$, consisting of three ATP-A (α) and three ATP-B (β) subunits, which are encoded by the chloroplast DNA (Rodermel & Bogorad, 1987). It has been reported that mutation or editing deficiency at a certain site results in a substitution of amino acid residue, impairing the assembly of chloroplast ATP

synthase (Liu et al., 2021). A large number of studies have suggested that when exposed to lower temperatures, resistant plants showed higher ATPase activity than susceptible ones, which provide extra energy in plant responses to low temperatures (Gilmore & Bjrkman, 1995; Schttler & Toth, 2014). Ultracytochemical localization studies on the effect of ATPase on low-temperature adaptability also provide direct evidence that chloroplast ATPase is closely related to the plant's tolerance to low temperatures (He, 2020). Our analysis showed that *K. obovata* had serine (S) at the 89th site in the ATP-A protein domain instead of tyrosine (Y), as in *K. candel* (Figure 8D). Molecular dynamics simulation indicated that the ATP-A protein of *K. obovata* had a higher binding ability with ADP (Figures 8E, F), which implied that its ATPase possessed higher efficiency in ATP formation. In addition, the gene *atpB* remained consistent within *K. candel* individuals but exhibited variations in SNP loci within the different populations of *K. obovata* (Figure 6). Hence, the mutation sites in the genes *atpA* and *atpB* of *K. obovata* populations might closely relate to their adaptation to the lower temperature in the northern part of the South China Sea, which endowed it with higher tolerance to lower temperature in the northward colonization of *K. obovata*.

The above results provide molecular proof for the divergence of *K. obovata* and *K. candel* in adaptation to different habitats. This also well explained the previous physiological experimental results obtained both from a common garden trial in Hong Kong (Maxwell, 1995) and in a lab chilling experiment conducted between these two *Kandelia* species (Short et al., 2021). In 2008, low air temperature in winter caused serious damage to several mangrove species, but *K. obovata* exhibited the highest cold tolerance among mangroves during extreme cold events in China (Chen et al., 2010), which could be more ecophysiological proof to support our conclusions gained from the cp genomes.

5 Conclusions

The main finding reported here first provides the whole cp genome of *K. candel*, and comparative analyses of the whole cp genomes were conducted between *K. candel* and *K. obovata*. Although the cp genome is regarded as having a highly conservative nature and a slow evolutionary rate, interspecific differences exist between *K. candel* and *K. obovata* in adaptation to differential environments geographically separated by the South China Sea, including the sizes of cp genomes, codon usages, repeat sequences, and indels of genes. Significant interspecific differences were found in *ndhD* and *atpA*, which are involved in photosynthetic electron transport and ATP formation, implying that energy generation plays a pivotal role in their ability to adapt to different temperatures and light, two main geographical environmental factors. Future comparative studies on photosynthetic electron transport and photophosphorylation caused by the genetic variations of *atpA* and *ndhD* in these species are needed to clarify their response mechanisms to varied light radiation and temperatures.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: GenBank database under accession numbers from ON969308 to ON969332.

Author contributions

XZ designed the research project. XX, YZ performed research. XX, QL, YS and XZ analyzed data. XX, XZ and YS wrote the manuscript. LC, WW, GC, WN, MI, PP, HZ prepared plant materials and contributed to scientific discussion and revision of the manuscript. All authors have read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (42076176), the Natural Science Foundation of Fujian Province of China (2022J01053), National Key Research and Development Program of China (2017YFC0506102), China-ASEAN Maritime Collaboration Fund, and Fieldwork Funds for graduate students of Xiamen University (2022FG022), and the Scientific and Technological Research Project for Social Welfare of Zhongshan City of China (2019B2005).

Acknowledgments

We thank Prof. Zhiliang Ji from Xiamen University for his support in protein structure analysis. We also thank PhD students Xiaofang Shi, Xiaobao Pan, Xiaoxuan Gu, Qiulian Lin, Jing Li and Dr. Anek Sophon, Dr. Ajcharaporn Piumsomboon, Mr. Shunyang Chen, Mr. Bingpeng Xing for their assistance in the experiments and sampling. In addition, we would like to give our thanks to the reviewers for their constructive suggestions in improving the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1075353/full#supplementary-material

References

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19 (9), 1655–1664. doi: 10.1101/gr.094052.109

Beier, S., Thiel, T., Münch, T., Scholz, U., and Mascher, M. (2017). MISA-web: A web server for microsatellite prediction. *Bioinf. (Oxford England)* 33 (16), 2583–2585. doi: 10.1093/bioinformatics/btx198

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* 27 (2), 573–580. doi: 10.1093/nar/27.2.573

Biju, V. C., Vijayan, S., Rajan, V. S., Sasi, A., Janardhanan, A., and Nair, A. S. (2019). The complete chloroplast genome of trichopus zeylanicus, and phylogenetic analysis with dioscoreales. *Plant Genome* 12 (3), 1–11. doi: 10.3835/plantgenome2019.04.0032

Birky, C. W. (2001). The inheritance of genes in mitochondria and chloroplasts: Laws, mechanisms, and models. *Annu. Rev. Genet.* 35, 125–148. doi: 10.1146/annurev.genet.35.102401.090231

Bobik, K., and Burch-Smith, T. M. (2015). Chloroplast signaling within, between and beyond cells. *Front. Plant Sci.* 6 (781). doi: 10.3389/fpls.2015.00781

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinf. (Oxford England)* 30 (15), 2114–2120. doi: 10.1093/bioinformatics/btu170

Braukmann, T. W. A., Kuzmina, M., and Stefanovi, S. (2009). Loss of all plastid ndh genes in gnetales and conifers: Extent and evolutionary significance for the seed plant phylogeny. *Curr. Genet.* 55 (3), 323–337. doi: 10.1007/s00294-009-0249-7

Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G. V., et al. (2021). RCSB protein data bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.* 49 (D1), D437–D451. doi: 10.1093/nar/gkaa1038

Casano, L. M., Zapata, J. M., Martín, M., and Sabater, B. (2000). Chlororespiration and poising of cyclic electron transport. Plastoquinone as electron transporter between thylakoid NADH dehydrogenase and peroxidase. *J. Biol. Chem.* 275 (2), 942–948. doi: 10.1074/jbc.275.2.942

Chen, F., Mackey, A. J., Stoeckert, C. J., and Roos, D. S. (2006). OrthoMCL-DB: Querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* 34 (1), D363–D368. doi: 10.1093/nar/gkj123

Chen, C., Rui, X., Hao, C., and He, Y. (2018). Tbtools, a toolkit for biologists integrating various hts-data handling tools with a user-friendly interface. *Cold Spring Harbor Lab.* 8. doi: 10.1101/289660

Chen, L. Z., Wang, W. Q., Zhang, Y. H., Huang, L., and Lin, G. H. (2010). Damage to mangroves from extreme cold in early 2008 in southern China. *Chin. J. Plant Ecol.* 34 (2), 186–194. doi: 10.1016/S0140-6736(00)30105-2

Chen, D. Q., Xiang, S., Liu, Z. J., and Zou, S. Q. (2019). The complete chloroplast genome sequence of kandelia obovata (Rhizophoraceae). *Mitochondrial DNA Part B* 4, 2, 3494–3495. doi: 10.1080/23802359.2019.1674745

Chiang, T. Y., Chiang, Y. C., Chen, Y. J., Chou, C. H., Havanond, S., Hong, T. N., et al. (2001). Phylogeography of kandelia candel in East Asiatic mangroves based on nucleotide variation of chloroplast and mitochondrial DNAs. *Mol. Ecol.* 10 (11), 2697–2710. doi: 10.1046/j.0962-1083.2001.01399.x

Cui, G. X., Wang, C. M., Wei, X. X., Wang, H. B., Wang, X. L., Zhu, X. Q., et al. (2021). Complete chloroplast genome of hordeum brevisubulatum: Genome organization, synonymous codon usage, phylogenetic relationships, and comparative structure analysis. *PloS One* 16 (12), e0261196. doi: 10.1371/journal.pone.0261196

Cvetkovska, M., Szyszka-Mroz, B., Possmayer, M., Pittock, P., Lajoie, G., Smith, D. R., et al. (2018). Characterization of photosynthetic ferredoxin from the Antarctic alga chlamydomonas sp. UWO241 reveals novel features of cold adaptation. *New Phytol.* 219 (2), 836–875. doi: 10.1111/nph.15194

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinf. (Oxford England)* 27 (15), 2156–2158. doi: 10.1093/bioinformatics/btr330

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

Darling, A. C. E., Mau, B., Blattner, F. R., and Perna, N. T. (2004). Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14 (7), 1394–1403. doi: 10.1101/gr.2289704

Das, A. B., Basak, U. C., and Das, P. (1995). Karyotype diversity and genomic variability in some Indian tree mangroves. *Caryologia* 48 (3–4), 319–328. doi: 10.1080/00087114.1995.10797341

Dierckxsens, N., Mardulyn, P., and Smits, G. (2020). Unraveling heteroplasmy patterns with NOVOPlasty. *NAR Genomics Bioinf.* 2 (1), lqz011. doi: 10.1093/nargab/lqz011

Doyle, J. (1991). DNA Protocols for plants: CTAB total DNA isolation. In *Molecular techniques in taxonomy*. (Springer : Berlin Heidelberg). 283–293.

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32 (5), 1792–1797. doi: 10.1093/nar/gkh340

Endo, T., Shikanai, T., Takabayashi, A., Asada, K., and Sato, F. (1999). The role of chloroplastic NAD(P)H dehydrogenase in photoprotection. *FEBS Lett.* 457 (1), 5–8. doi: 10.1016/S0014-5793(99)00989-8

Enoki, T., Ueda, M., Nanki, D., Suwa, R., and Hagihara, A. (2009). Distribution and stem growth patterns of mangrove species along the nakara river in iriomote island, southwestern Japan. *J. For. Res.* 14 (1), 51–54. doi: 10.1007/s10310-008-0094-4

Fang, J. P., Lin, A. T., Yuan, X., Chen, Y. Q., He, W. J., Huang, Y. J., et al. (2020). The complete chloroplast genome of isochrysis galbana and comparison with related haptophyte species. *Algal Res.* 50, 101989. doi: 10.1016/j.algal.2020.101989

Feng, C., You, H. M., Tan, F. L., Han, J. L., Yu, X. X., You, W. B., et al. (2022). Methane contributions of different components of kandelia candel–soil system under nitrogen supplementation. *Forests* 13 (2), 318. doi: 10.3390/f13020318

Fields, P. A., Dong, Y., Meng, X., and Somero, G. N. (2015). Adaptations of protein structure and function to temperature: there is more than one way to 'skin a cat'. *J. Exp. Biol.* 218 (Pt 12), 1801–1811. doi: 10.1242/jeb.114298

Geng, Q., Lian, C., Goto, S., Tao, J., Kimura, M., Islam, M. S., et al. (2008). Mating system, pollen and propagule dispersal, and spatial genetic structure in a high-density population of the mangrove tree kandelia candel. *Mol. Ecol.* 17 (21), 4724–4739. doi: 10.1111/j.1365-294X.2008.03948.x

Giang, L. H., Geada, G. L., Hong, P. N., Tuan, M. S., Lien, N. T. H., Ikeda, S., et al. (2006). Genetic variation of two mangrove species in kandelia (Rhizophoraceae) in Vietnam and surrounding area revealed by microsatellite markers. *Int. J. Plant Sci.* 167 (2), 291–298. doi: 10.1086/499611

Gilmore, A. M., and Bjrkman, O. (1995). Temperature-sensitive coupling and uncoupling of ATPase-mediated, nonradiative energy dissipation: Similarities between chloroplasts and leaves. *Planta* 197 (4), 646–654. doi: 10.1007/BF00191573

Greiner, S., Lehwark, P., and Bock, R. (2019). OrganellarGenomeDRAW (OGDRAW) version 1.3.1: Expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* 47 (W1), W59–W64. doi: 10.1093/nar/gkz238

Groth, G., and Pohl, E. (2001). The structure of the chloroplast F1-ATPase at 3.2 a resolution. *J. Biol. Chem.* 276, 1345–1352. doi: 10.1074/jbc.M008015200

Gupta, P. K., Balyan, H. S., Edwards, K. J., Isaac, P., Korzun, V., Rder, M., et al. (2022). Genetic mapping of 66 new microsatellite (SSR) loci in bread wheat. TAG. theoretical and applied genetics. *Theor. Appl. Genet.* 105 (2–3), 413–422. doi: 10.1007/s00122-002-0865-9

Hahn, A., Vonck, J., Mills, D. J., Meier, T., and Kühlbrandt, W. (2018). Structure, mechanism, and regulation of the chloroplast ATP synthase. *Science* 360 (6389), eaat4318. doi: 10.1126/science.aat4318

He, J. Y. (2020). Study on the effect of ATPase on the low-temperature adaptability of galinsoga parviflora cav. by ultracytochemical localization. *J. Electron Microscopy* 39 (03), 307–312. doi: 10.3969/j.issn.1000-6281.2020.03.013

Hou, D. (1958). "Flora malesiana," in *Rhizophoraceae* Ed. C. G. G. J. van Steenis (Djakarta: Noordhoff-ko lffN.V.), pp 429–493.

Jin, J. J., Yu, W. B., Yang, J. B., Song, Y., dePamphilis, C. W., Yi, T. S., et al. (2020). GetOrganelle: A fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. *Genome Biol.* 21 (1), 241. doi: 10.1186/s13059-020-02154-5

Joet, T., Cournac, L., Peltier, G., and Havaux, M. (2002). Cyclic electron flow around photosystem I in C(3) plants. *In vivo* control by the redox state of chloroplasts and involvement of the NADH-dehydrogenase complex. *Plant Physiol.* 128 (2), 760–769. doi: 10.1104/pp.010775

Johnson, M., Mark, J., Irena, Z., Yan, R., Yuri, M., Scott, M. G., et al. (2008). NCBI BLAST: a better web interface. *Nucleic Acids Res.* 36, W5–W9. doi: 10.1093/nar/gkn201

Juncosa, A. M., and Tomlinson, P. B. (1988). A historical and taxonomic synopsis of rhizophoraceae and anisophylleaceae. *Ann. Missouri Botanical Garden* 75), 1278–1295. doi: 10.2307/2399286

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* 30 (4), 772–780. doi: 10.1093/molbev/mst010

Khakhlova, O., and Bock, R. (2006). Elimination of deleterious mutations in plastid genomes by gene conversion. *Plant Journal: For Cell Mol. Biol.* 46 (1), 85–94. doi: 10.1111/j.1365-313X.2006.02673.x

Kim, K. A., and Cheon, K. S. (2021). Complete chloroplast genome sequence of adenophora racemosa (Campanulaceae): Comparative analysis with congeneric species. *PloS One* 16 (3), e0248788. doi: 10.1371/journal.pone.0248788

Kong, B. L. H., Park, H. S., Lau, T. W. D., Lin, Z., Yang, T. J., and Shaw, P. C. (2021). Comparative analysis and phylogenetic investigation of Hong Kong ilex chloroplast genomes. *Sci. Rep.* 11 (1), 5153. doi: 10.1038/s41598-021-84705-9

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35 (6), 1547–1549. doi: 10.1093/molbev/msy096

Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. (2001). REPuter: The manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* 29 (22), 4633–4642. doi: 10.1093/nar/29.22.4633

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nat. Methods* 9 (4), 357–359. doi: 10.1038/nmeth.1923

Letunic, I., Khedkar, S., and Bork, P. (2021). SMART: Recent updates, new developments and status in 2020. *Nucleic Acids Res.* 49 (D1), D458–D460. doi: 10.1093/nar/gkaa937

Liao, B., and Zhang, Q. (2014). Area, distribution and species composition of mangroves in China. *Chin. J. Wetland Sci.* 4 (12), 435–440. doi: 10.13248/j.cnki.wetlandsci.2014.04.005

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 25 (14), 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, Q. H., He, Z. H., and Mi, H. L. (2013). The research progress of chloroplast NAD(P)H dehydrogenase (NDH) complex. *Chin. J. Plant Physiol. J.* 49 (5), 401–409. doi: 10.13592/j.cnki.ppj.2013.05.008

Li, M. S., and Lee, S. Y. (1997). Mangroves of China: A brief review. *For. Ecol. Manage.* 96 (3), 241–259. doi: 10.1016/S0378-1127(97)00054-6

Lin, P. (1999). *Mangrove ecosystem in China* (Beijing: Science Press).

Liu, X. Y., Jiang, R. C., Wang, Y., Tang, J. J., and Tan, B. C. (2021). ZmPPR26, a DYW-type pentatricopeptide repeat protein, is required for c-to-U RNA editing at atpA-1148 in maize chloroplasts. *J. Exp. Bot.* 72 (13), 4809-4821. doi: 10.1093/jxb/erab185

Li, J. Q., Xu, H. F., Ye, L. Z., Gu, J. F., and Wang, R. Q. (2001). Introduction and afforestation technique of Kandelia candel to the North. *Zhejiang Forestry Science and Technology* 6, 51–53. Available at: https://www.docin.com/p-1526668739.html

Luo, C., Huang, W. L., Sun, H. Y., Yer, H. Y., Li, X. Y., Li, Y., et al. (2021). Comparative chloroplast genome analysis of impatiens species (Balsaminaceae) in the karst area of China: Insights into genome evolution and phylogenomic implications. *BMC Genomics* 22 (1), 571. doi: 10.1186/s12864-021-07807-8

Martín, M., and Sabater, B. (2010). Plastid ndh genes in plant evolution. *Plant Physiol. Biochem.* 48 (8), 636–645. doi: 10.1016/j.plaphy.2010.04.009

Maxwell, G. S. (1995). Ecogeographic variation in from Brunei, Hong Kong and Thailand. *Asia-Pacific Symposium Mangrove Ecosyst.* 106, 59–65. doi: 10.1007/978-94-011-0289-6_8

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20 (9), 1297–1303. doi: 10.1101/gr.107524.110

Naskar, K. R., and Mandal, R. N. (1999). Ecology and biodiversity of Indian mangroves. *global status.* (Delhi: Daya Publishing House 2, 397–400.

Nathan, S., Andreas, P. E., Alexandra, C., Sereina, R., Moritz, W., Alan, E. M., et al. (2011). Definition and testing of the GROMOS force-field versions 54A7 and 54B7. *Eur. Biophysics J.* 40, 843–856. doi: 10.1007/s00249-011-0700-9

Nelson, N., and Ben-Shem, A. (2004). The complex architecture of oxygenic photosynthesis. *Nat. Rev. Mol. Cell Biol.* 5 (12), 971. doi: 10.1038/nrm1525

Peredo, E. L., King, U. M., and Les, D. H. (2013). The plastid genome of najas flexilis: Adaptation to submersed environments is accompanied by the complete loss of the NDH complex in an aquatic angiosperm. *PloS One* 8 (7), e68591. doi: 10.1371/journal.pone.0068591

Powell, W., Machray, G. C., and Provan, J. (1996). Polymorphism revealed by simple sequence repeats. *Trends Plant Sci.* 1 (7), 215–222. doi: 10.1016/1360-1385(96)86898-1

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: The European molecular biology open software suite. *Trends Genetics: TIG* 16 (6), 276–277. doi: 10.1016/s0168-9525(00)02024-2

Rodermel, S. R., and Bogorad, L. (1987). Molecular evolution and nucleotide sequences of the maize plastid genes for the alpha subunit of CF1 (atpA) and the proteolipid subunit of CF0 (atpH). *Genetics* 116 (1), 127–139. doi: 10.1093/genetics/116.1.127

Rousseau Gueutin, M., Bellot, S., Martin, G. E., Boutte, J., Chelaifa, H., Lima, O., et al. (2015). The chloroplast genome of the hexaploid spartina maritima (Poaceae, chloridoideae): Comparative analyses and molecular dating. *Mol. Phylogenet. Evol.* 93, 5–16. doi: 10.1016/j.ympev.2015.06.013

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., et al. (2017). DnaSP 6: DNA sequence polymorphism analysis of Large data sets. *Mol. Biol. Evol.* 34 (12), 3299–3302. doi: 10.1093/molbev/msx248

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., and Tabata, S. (1999). Complete structure of the chloroplast genome of arabidopsis thaliana. *DNA Research: Int. J. Rapid Publ. Rep. Genes Genomes* 6 (5), 283–290. doi: 10.1093/dnares/6.5.283

Schttler, M. A., and Toth, S. Z. (2014). Photosynthetic complex stoichiometry dynamics in higher plants: Environmental acclimation and photosynthetic flux control. *Front. Plant Sci.* 5 (188). doi: 10.3389/fpls.2014.00188

Schuller, J. M., Birrell, J. A., Tanaka, H., Konuma, T., Wulfhorest, H., Cox, N., et al. (2019). Structural adaptations of photosynthetic complex I enable ferredoxin-dependent electron transfer. *Science* 363, 257–260. doi: 10.1126/science.aau3613

Shen, L. L., Tang, K. L., Wang, W. D., Wang, C., Wu, H., Mao, Z. Y., et al. (2021). Architecture of the chloroplast PSI-NDH supercomplex in hordeum vulgare. *Nature* 601, 649–654. doi: 10.1038/s41586-021-04277-6

Sheue, C. R., Liu, H. Y., and Yong, J. W. H. (2003). *Kandelia obovata* (Rhizophoraceae), a new mangrove species from Eastern Asia. *Taxon* 52, 2, 287–294. doi: 10.2307/3647398

Shi, L. C., Chen, H. M., Jiang, M., Wang, L. Q., Wu, X., Huang, L. F., et al. (2019). CPGAVAS2, an integrated plastome sequence annotator and analyzer. *Nucleic Acids Res.* 47 (W1), W65–W73. doi: 10.1093/nar/gkz345

Shikanai, T. (2016). Chloroplast NDH: A different enzyme with a structure similar to that of respiratory NADH dehydrogenase. *Biochim. Biophys. Acta (BBA) - Bioenergetics* 1857 (7), 1015–1022. doi: 10.1016/j.bbabio.2015.10.013

Shikanai, T. (2020). Regulation of photosynthesis by cyclic electron transport around photosystem I. *Adv. Botanical Res* 96, 177–204. doi: 10.1016/bs.abr.2020.07.005

Short, A. W., Chen, R., and Wee, A. (2021). Comparison between parapatric mangrove sister species revealed higher photochemical efficiency in subtropical than tropical coastal vegetation under chilling stress. *Aquat. Bot.* 168, 103323. doi: 10.1016/j.aquabot.2020.103323

Silva, S. R., Diaz, Y. C. A., Penha, H. A., Pinheiro, D. G., Fernandes, C. C., Miranda, V. F. O., et al. (2016). The chloroplast genome of utricularia reniformis sheds light on the evolution of the ndh gene complex of terrestrial carnivorous plants from the lentibulariaceae family. *PloS One* 11 (10):e0165176. doi: 10.1371/journal.pone.0165176

Steele, K. A., Price, A. H., Shashidhar, H. E., and Witcombe, J. R. (2006). Marker-assisted selection to introgress rice QTLs controlling root traits into an Indian upland rice variety. TAG. theoretical and applied genetics. *Theoretische Und Angewandte Genetik* 112 (2), 208–221. doi: 10.1007/s00122-005-0110-4

Sun, Z. T., Liu, Q., Qu, G., Feng, Y., and Reetz, M. T. (2019). Utility of b-factors in protein science: Interpreting rigidity, flexibility, and internal motion and engineering thermostability. *Chem. Rev.* 119 (3), 1626–1665. doi: 10.1021/acs.chemrev.8b00290

Takeuchi, T., Sugaya, T., Kanazashi, A., and Katsuta, Y. M. (2001). Genetic diversity of kandelia candel and bruguiera gymnorrhiza in the southwest islands, Japan. *J. For. Res.* 6 (3), 157–162. doi: 10.1007/BF02767087

Tian, J., Wang, P., and Ning F., W. U. (2012). Recent advances in the rational design to improve the protein thermostability. *Chin. J. Curr. Biotechnol.* 2 (04), 233–239. doi: 10.3969/j.issn.2095-2341.2012.04.01

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., et al. (2017). GeSeq–versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* 45 (W1), W6–W11. doi: 10.1093/nar/gkx391

Tomlinson, P. B. (1986). *The botany of mangroves* (Cambridge University Press).

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. C. (2005). GROMACS: Fast, flexible, and free. *J. Comput. Chem.* 26 (16), 1701–1718. doi: 10.1002/jcc.20291

Wang, N. J., Chen, S. F., Xie, L., Wang, L., Feng, Y. Y., Lv, T., et al. (2022). The complete chloroplast genomes of three hamamelidaceae species: Comparative and phylogenetic analyses. *Ecol. Evol.* 12 (2), e8637. doi: 10.1002/ece3.8637

Wang, C. X., Liu, J. J., Su, Y., Li, M. L., Xie, X. Y., and Su, J. J. (2021). Complete chloroplast genome sequence of sonchus brachyotus helps to elucidate evolutionary relationships with related species of asteraceae. *BioMed. Res. Int.* 2021, 9410496. doi: 10.1155/2021/9410496

Wang, W. Q., and Wang, M. (2007). *The mangroves of China* (Beijing: Scientific Press).

Wang, D. P., Zhang, Y. B., Zhang, Z., Zhu, J., and Yu, J. (2010). KaKs_Calculator 2.0: A toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinf.* 8 (1), 77–80. doi: 10.1016/S1672-0229(10)60008-3

Wei, F., Tang, D. F., Wei, K. H., Qin, F., Li, L. X., Lin, Y., et al. (2020). The complete chloroplast genome sequence of the medicinal plant sophora tonkinensis. *Sci. Rep.* 10 (1), 12473. doi: 10.1038/s41598-020-69549-z

Yang, L. C., Xiong, F., Xiao, Y. M., Li, J. J., Chen, C., Zhou, G. Y., et al. (2019a). The complete chloroplast genome of swertia tetraptera and phylogenetic analysis. *Mitochondrial DNA. Part B, Resources* 5 (1), 164–165. doi: 10.1080/23802359.2019.1698368

Yang, Y., Zhang, Y., Chen, Y. K., Gul, J. M., Zhang, J. W., Liu, Q., et al. (2019b). Complete chloroplast genome sequence of the mangrove species kandelia obovata and comparative analyses with related species. *PeerJ* 7, e7713. doi: 10.7717/peerj.7713

Yoshioka, H., Kondo, K., Segawa, M., Nehira, K., and Maeda, S. (1984). Karyomorphological studies in five species of mangrove genera in the rhizophoraceae. *Kromosomo* 2 (35–36), 1111–1116.

Yukawa, M., Tsudzuki, T., and Sugiura, M. (2005). The 2005 version of the chloroplast DNA sequence from tobacco (Nicotiana tabacum). *Plant Mol. Biol. Rep.* 23 (4), 359–365. doi: 10.1007/BF02788884

Zalapa, J. E., Cuevas, H., Zhu, H., Steffan, S., Senalik, D., Zeldin, E., et al. (2012). Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. *Am. J. Bot.* 99 (2), 193–208. doi: 10.3732/ajb.1100394

Zhang, F. Q., Wang, Y. S., Sun, C. C., Lou, Z. P., and Dong, J. D. (2012a). A novel metallothionein gene from a mangrove plant kandelia candel. *Ecotoxicology* 21 (6), 1633–1641. doi: 10.1007/s10646-012-0952-x

Zhang, Z., Xiao, J. F., Wu, J. Y., Zhang, H. Y., Liu, G. M., Wang, X. M., et al. (2012b). ParaAT: A parallel tool for constructing multiple protein-coding DNA alignments. *Biochem. Biophys. Res. Commun.* 419 (4), 779–781. doi: 10.1016/j.bbrc.2012.02.101

Zhu, X. Y., Chen, G. C., and Zhang, C. L. (2001). Photosynthetic electron transport, photophosphorylation, and antioxidants in two ecotypes of reed (Phragmites communis trin.) from different habitats. *Photosynthetica* 39 (2), 183–189. doi: 10.1023/A:1013766722604

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 06 January 2023
DOI 10.3389/fpls.2022.1099856

OPEN ACCESS

EDITED BY
Linchun Shi,
Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences and Peking Union Medical College, China

REVIEWED BY
Wenjun Zhu,
Wuhan Polytechnic University, China
Wei Sun,
China Academy of Chinese Medical Sciences, China

*CORRESPONDENCE
Yinglin Liu
lyldaliedu@163.com
Ping Zhou
zhouping725@126.com

SPECIALTY SECTION
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

RECEIVED 16 November 2022
ACCEPTED 13 December 2022
PUBLISHED 06 January 2023

CITATION
Fang H, Dai G, Liao B, Zhou P and Liu Y (2023) Application of chloroplast genome in the identification of *Phyllanthus urinaria* and its common adulterants.
Front. Plant Sci. 13:1099856.
doi: 10.3389/fpls.2022.1099856

Application of chloroplast genome in the identification of *Phyllanthus urinaria* and its common adulterants

Hui Fang, Guona Dai, Binbin Liao, Ping Zhou* and Yinglin Liu*

College of Pharmaceutical Science, Dali University, Dali, China

Background: *Phyllanthus urinaria* L. is extensively used as ethnopharmacological material in China. In the local marketplace, this medicine can be accidentally contaminated, deliberately substituted, or mixed with other related species. The contaminants in herbal products are a threat to consumer safety. Due to the scarcity of genetic information on *Phyllanthus* plants, more molecular markers are needed to avoid misidentification.

Methods: In this study, the complete chloroplast genome of nine species of the genus *Phyllanthus* was *de novo* assembled and characterized.

Results: This study revealed that all of these species exhibited a conserved quadripartite structure, which includes a large single copy (LSC) region and small single copy (SSC) region, and two copies of inverted repeat regions (IRa and IRb), which separate the LSC and SSC regions. And the genome structure, codon usage, and repeat sequences were highly conserved and showed similarities among the nine species. Three highly variable regions (*trnS-GCU-trnG-UCC*, *trnT-UGU-trnL-UAA*, and *petA-psbJ*) might be helpful as potential molecular markers for identifying P. urinaria and its contaminants. In addition, the molecular clock analysis results showed that the divergence time of the genus *Phyllanthus* might occur at ~ 48.72 Ma.

Conclusion: This study provides valuable information for further species identification, evolution, and phylogenetic research of *Phyllanthus*.

KEYWORDS

***Phyllanthus urinaria*, chloroplast genome, species identification, molecular marker, phylogenetic**

1 Introduction

Phyllanthus urinaria L. belongs to the family Euphorbiaceae, is listed in the dictionary of Chinese ethnic medicine, and the Chinese name is "Yexiazhu" (Guo et al., 2017). The herbal *P. urinaria* has crucial medicinal value in anti-diarrheal, anti-inflammatory, jaundice, diabetes, malaria, hepatitis B, and liver diseases (Chudapongse et al., 2010; Geethangili and Ding, 2018). The previous survey has revealed that *P. urinaria* is generally contaminated with other common adulterants, such as *P. acidus* (L.) Skeel, *P. amarus* Schumacher & Thonning, *P. reticulatus* Poir., *P. niruri* L., *P. emblica* L., *P. pulcher* Well. ex Muell. Arg., and *P. debilis* Klein ex Willd. (Manissorn et al., 2010; Srirama et al., 2010; Kiran et al., 2021). These adulterants are usually of poor quality, and some might even be toxic (Adedapo et al., 2005; Geng et al., 2021). As the morphology of these species are interchangeable, similar, and indistinguishable, the identification of these species remains controversial, which may impair their clinical safety and efficacy (Sarin et al., 2013; Kiran et al., 2021). Therefore, it is essential to develop a method for accurately identifying *P. urinaria* and its common contaminants.

With the rapid development of molecular technology, molecular identification has made significant progress in Chinese medicine, especially molecular markers, a technique that involves sequencing specific sections of the genome to identify differences between individuals of different species or populations (Xiong et al., 2018). Recent studies have revealed high levels of genetic diversity and a lack of phylogenetic resolution within species of *Phyllanthus* (Pruesapan et al., 2012; Bouman et al., 2021). Universal DNA barcodes, such as ITS, *psbA-trnH*, *trnL*, *psbK-psbI*, *rpoC1*, and *trnL-trnF*, have been used to identify *P. urinaria* and its adulterants (Manissorn et al., 2010; Srirama et al., 2010; Inglis et al., 2018; Kiran et al., 2021). However, some common adulterants were not included in these investigations, and there are inherent limitations to single-locus DNA barcodes (Heinze, 2007; Li et al., 2015). Therefore, more scientific and accurate identification methods must be developed. The chloroplast (cp) is an essential organelle that plays a crucial role in plant photosynthesis and several other critical biochemical processes (Neuhaus and Emes, 2000). Compared with the traditional DNA fragments, the cp genome was relatively conserved and slightly varied, which has been applied to many research fields, such as species identification and the development of molecular markers (Abdullah et al., 2020; Li et al., 2022; Wang et al., 2022). The method has been widely used for identifying *Paris*, *Polygonatum*, *Vicatia*, and their adulterants (Guan et al., 2022; Jiang et al., 2022; Wang et al., 2022). Recently, although the complete cp genomes of Phyllanthaceae species have been reported and the high-resolution phylogenetic tree was reconstructed (Rehman et al., 2021), the purpose of this study was to clarify the genome evolution in Phyllanthaceae and identify the polymorphic loci for phylogenetic inference. To our knowledge, no reports use cp genomes to compare *P. urinaria* with its common adulterants.

Our study aims to: (i) contribute new fully-sequenced cp genomes of *Phyllanthus* and improve the understanding of the overall structure of these genomes, (ii) perform comparative analyses and elucidate the phylogenetic evolution of the *Phyllanthus*, and (iii) screen potential molecular markers to distinguish *P. urinaria* from its contaminants. In the current work, the complete cp genomes of nine *Phyllanthus* species were sequenced, *de novo* assembled, and annotated. These genomes were then used in a comparative analysis of genome structure and evolution relationships. This research expands the genomic resources available for *Phyllanthus* and provides valuable information support for the phylogenetic analysis and identification of the *Phyllanthus*, as well as for the safe applications of *P. urinaria*.

2 Material and methods

2.1 Plant and DNA sources

The fresh and healthy leaves for nine species of *P. acidus*, *P. amarus*, *P. reticulatus*, *P. urinaria*, *P. niruri*, *P. niruri* subsp. *lathyroides*, *P. emblica*, *P. pulcher*, and *P. franchetianus* were collected from Dali and Xishuangbanna, Yunnan Province, China. The detailed information per sample is available in Supplementary Table 1. The samples were identified following the taxonomic key and external morphology diagnosis proposed by related literature (Webster and Carpenter, 2008). The voucher specimens were preserved at the herbarium of Dali University. The fresh leaf of nine species was frozen in liquid nitrogen and stored in a 4°C refrigerator for DNA extraction. Total DNA was extracted using a modified cetyl trimethyl ammonium bromide (CTAB) procedure (Allen et al., 2006). DNA quality and quantity were assessed using a NanoDrop spectrophotometer (ND-2000; Thermo Fisher Scientific, USA) and agarose gel electrophoresis.

2.2 DNA sequencing, assembly and annotation

Purified high-quality genomic DNA was broken into short fragments of approximately 350 bp, and paired-end (PE) libraries were constructed by adding A-tails, PCR amplification, and other steps, followed by sequencing in 150 bp paired-end mode on an Illumina NovaSeq 6000 platform. The high-quality reads were assembled using GetOrganelle v1.7.5 (Jin et al., 2020) and then annotated by cpGAVAS2 (http://47.96.249.172:16019/analyzer/annotate) and (GeSeq, RRID : SCR_017336) (https://chlorobox.mpimp-golm.mpg.de/geseq.html) (Tillich et al., 2017; Shi et al., 2019). The annotations of tRNA genes were confirmed by using

(tRNAscan-SE v.2.03, RRID : SCR_010835) (http://lowelab.ucsc.edu/tRNAscan-SE/) (Schattner et al., 2005). Annotated cp genomes sequences were submitted to GenBank and are available under accession numbers OP009343-OP009351 (Table 1). Fully annotated cp genome circular diagrams were drawn by OrganellarGenomeDRAW (OGDRAW, RRID : SCR_017337) (https://chlorobox.mpimp-golm.mpg.de/OGDraw.html) (Greiner et al., 2019).

2.3 Genome structure and comparisons analysis

Forward (F), palindromic (P), reverse (R), and complementary (C) were identified using the REPuter (https://bibiserv.cebitec.uni-bielefeld.de/reputer/) tool (Kurtz et al., 2001). The criteria for identifying repeats include a minimum repetition size of 30 bp and a 90% similarity between repeat pairs, calculated by assigning a value of 3 to the altered sequence. In addition, (MISA, RRID : SCR_010765) (http://pgrc.ipk-gatersleben.de/misa/) software was used to identify simple sequence repeats (SSRs) (Beier et al., 2017). We followed conventional standards for identifying chloroplast and mitochondrial SSRs, including a minimum stretch of 10 for mono-, 5 for di-, 4 for tri-, and 3 for tetra-, penta- and hexanucleotide repeats and a minimum distance of 100 bp between compound SSRs. Relative synonymous codon usage (RSCU) was analyzed by CodonW v.1.4.2 (Sharp et al., 1986). Also, Tbtools v1.098761 used a heatmap to show the values of RSCU (Chen et al., 2020).

For comparative analysis of genes, tRNA, repeat content, genome size, and GC content were assessed by (Geneious v.2022.2.1, RRID : SCR_010519) (http://www.geneious.com/) software (Kearse et al., 2012). The software mVISTA (https://genome.lbl.gov/vista/index.shtml) (Frazer et al., 2004) in Shuffle-LAGAN mode (Brudno et al., 2003) was used to compare the nine *de novo* cp genome sequences, *P. amarus* (GenBank OP009344) was used as the reference genome. IRscope (https://irscope.shinyapps.io/irapp/) was used to analyze inverted repeat region contraction and expansion at the junctions of cp genomes (Amiryousefi et al., 2018). The cp genomes were aligned in (MAFFT, RRID : SCR_011811) (https://mafft.cbrc.jp/alignment/server/). Additionally, the nucleotide variability (Pi) across the cp genome sequences was performed in (DnaSP v.6.12.03, RRID : SCR_003067) (http://www.ub.edu/dnasp/) (Rozas et al., 2017), with a window length of 600 bp and step size of 200 bp. A value of Pi higher than 0.05 was recommended as mutational hotspots (Ren et al., 2022).

TABLE 1 Cp genomes features of nine species of *Phyllanthus*.

Genome features	*P. acidus*	*P. amarus*	*P. reticulatus*	*P. urinaria*	*P. niruri*	*P. niruri* subsp. *lathyroides*	*P. emblica*	*P. pulcher*	*P. franchetianus*
Genome size (bp)	156,331	155,790	156,610	153,850	155,900	143,563	155,841	155,589	155,598
LSC size (bp)	85,807	85,185	85,868	83,714	85,307	91,305	85,721	85,533	85,533
SSC size (bp)	19,262	19,015	19,182	18,780	19,015	18,986	18,950	18790	18,799
IRa/IRb size (bp)	25,631	25,795	25,780	25,678	25,789	16,771	25,585	25,633	25,633
Total GC content (%)	36.9	36.6	36.6	36.9	36.6	36.8	36.8	36.8	36.8
GC content in LSC (%)	34.6	34.3	34.3	34.5	34.2	34.9	34.5	34.4	34.4
GC content in S.S.C. (%)	30.6	30.0	30.2	30.6	30.0	30.1	30.4	30.9	30.9
GC content in IRa/IRb (%)	43.1	42.9	42.9	43.0	42.9	45.6	43.1	42.9	42.9
Number of genes	126	125	125	122	125	118	126	123	123
Protein-coding genes	82	81	82	79	81	75	82	79	79
tRNA genes	36	36	35	35	36	35	36	36	36
rRNA genes	8	8	8	8	8	8	8	8	8
Accession numbers in GenBank	OP009343	OP009344	OP009345	OP009346	OP009347	OP009348	OP009349	OP009350	OP009351

2.4 Phylogenetic analysis and divergence times analysis

For the phylogenetic analysis, 38 Euphorbiaceae taxa initially consisted of 29 species downloaded from NCBI (Table S2) and the 9 species presented here (Table 1). At the same time, two species, *Daphniphyllum oldhamii* (GenBank NC037883) and *D. macropodum* (GenBank MN496060) were selected as outgroups (Chase et al., 2016). A total of 40 cp genomes were aligned using MAFFT with the default parameters and trimmed using (TrimAl v.1.4, RRID : SCR_017334) (http://trimal.cgenomics.org/) with an automated option (Katoh and Standley, 2013). The best-fit model of nucleotide substitution was selected using ModelFinder (Kalyaanamoorthy et al., 2017) with the Bayesian information criterion as implemented in (IQ-tree v.1.6.12, RRID : SCR_017254) (http://www.iqtree.org/) (Nguyen et al., 2015). The alignment was also evaluated using bootstrap analysis on 1,000 in a maximum likelihood (ML) by IQ-tree v.1.6.12, with the following parameters: iqtree -s input -m TVM+F+R3 -bb 1000 -alrt 1000 -nt AUTO -o NC_037883, MN_496060. Besides, the neighbor-joining (NJ) tree was constructed using (MEGA X v.10.2.6, RRID : SCR_000667) (http://megasoftware.net/), and the bootstrap testing was performed with 1,000 repetitions (Kumar et al., 2018).

For analysis of divergence times, the molecular clock tree was constructed based on an ML tree using MEGA X v.10.2.6 (Kumar et al., 2018; Mello, 2018). The relevant divergence times were executed in the (TimeTree, RRID : SCR_021162) (http://www.timetree.org/) Resource (Kumar et al., 2017). Four calibration points were used to restrict each node: (F1) 110.9–121.0 Ma for the root node, (F2) 48.6–55.8 Ma for Phyllanthoideae stem age, (F3) 3.5–74.3 Ma for Acalyphoideae crown age, and (F4) 21.4–89 Ma for Euphorbioideae + Crotonoideae.

3 Results

3.1 Genome structure

The raw data of nine species were filtered to remove adapters and low-quality reads. Approximately 2.24–4.09 Gb data were obtained for each species. The cp genomes of these nine species are small circular DNA molecules with sizes in the range of 143,563 bp (*P. niruri* subsp. *lathyroides*) to 156,610 bp (*P. reticulatus*) (Figure 1), with the typical quadripartite structure of land plant cp genomes consisting of two inverted repeats (IRa and IRb) separated by large single copy (LSC) and small single copy (SSC) regions, respectively. The size of LSC ranged from 83,714 bp (*P. urinaria*) to 91,305 bp (*P. niruri* subsp. *lathyroides*), SSC ranged from 18,780 bp (*P. urinaria*) to 19,262 bp (*P. acidus*), and the size of each IR region ranged from 16,771 bp (*P. niruri* subsp. *lathyroides*) to 25,795 bp (*P. amarus*). Moreover, the GC content in the IR region (42.9%–45.6%) was higher than LSC (34.2%–34.9%) and SSC (30.0%–30.9%) (Table 1).

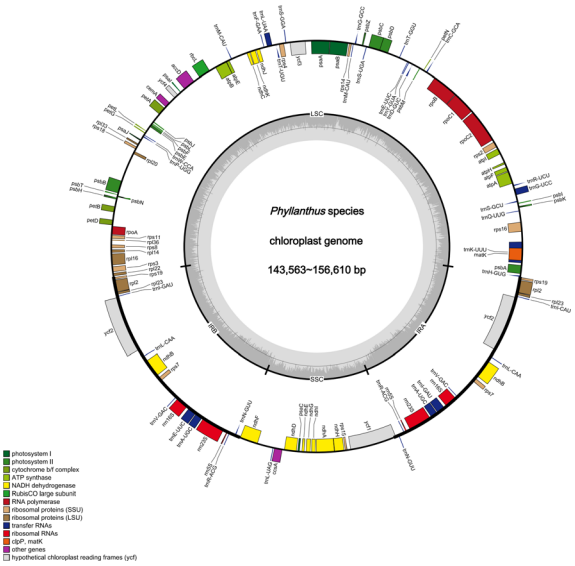


FIGURE 1
Cp genomes map of *Phyllanthus*. Genes inside and outside the circle are transcribed clockwise and counter-clockwise.

In addition, a total of 118–126 genes were identified, which comprised 75–82 protein-coding genes, 35–36 tRNAs, and 8 rRNAs (Table 1), whereas the number of genes varies in species due to IRs contraction and expansion. These genes were divided into three parts, of which 45 genes belong to photosynthesis-related genes (photosystem I, photosystem II, NADPH dehydrogenase, cytochrome b/f complex ATP synthase, and rubisco), 27 genes belong to self-replication (the large subunit of the ribosome, small subunit of the ribosome, and DNA dependent RNA polymerase), and the remaining genes belong to other genes (acetyl-CoA-carboxylase, c-type cytochrome synthesis genes, envelop membrane proteins, proteases, and maturase) (Table S3). Moreover, 17 genes each contained one intron, among them *rpl2* (×2), *ndhB* (×2), *trnI-GAU* (×2), and *trnA-UGC* (×2), which were located in the IR, and the genes (*trnK-UUU*, *rps16*, *trnG-UCC*, *rpoC1*, *ycf3*, *trnL-UAA*, *trnV-UAC*, and *clpP*) were located in the LSC, while the *ndhA* was only present in the SSC region. In addition, the *ycf3* and *clpP* each contain two introns (Table S3).

3.2 Repeat analysis

Repetitive sequences in cp genomes play a critical role in genome evolution and rearrangements. Analysis of oligonucleotide repeat revealed that the number of repeat types varied among the nine cp genomes and presented random permutations, and most repeat sequences were within 30–39 bp

(Figure S1). Meanwhile, the frequency of F and P repeats was greater than that of R and C repeats. The structural analysis of the repetition sequence is shown in Figure S1. The minimum number of repeats was found in *P. niruri* subsp. *lathyroides* (26), whereas the maximum was found in *P. niruri* (49).

SSRs, also known as microsatellites, consists of repeating units of 1–6 bp in length. 56, 98, 62, 80, 97, 54, 81, 69, and 69 SSRs were identified in *P. acidus*, *P. amarus*, *P. reticulatus*, *P. urinaria*, *P. niruri*, *P. niruri* subsp. *lathyroides*, *P. emblica*, *P. pulcher*, and *P. franchetianus*, respectively (Table S4). The number of SSRs showed the highest in *P. amarus* (98) and the lowest in *P. niruri* subsp. *lathyroides* (54). Most SSRs were found in LSC regions instead of SSC and IR regions (Figure S2). More than half of the SSRs (51.85%–66.67%) were mononucleotides with the A/T motif, followed by dinucleotides (16.25%–31.48%) with a predominant motif of AT/TA, trinucleotides (1.85%–6.25%) with a predominant motif of AAT/ATT, tetranucleotide repeats (1.61%–4.35%) with a predominant motif of AAAT/ATTT, pentanucleotides (0–1.61%), and hexanucleotides (0–1.25%) that only exist in the cp genome of *P. urinaria*.

3.3 Codon usage bias of cp genomes

The analyses of RSCU provide information about the encoding frequency of codons for an amino acid. There were 64 codons in the coding sequence of nine *Phyllanthus* species genes, among which 61 codons encoded 20 amino acids, and the other three codons (UAA, UAG, and UGA) were stop codons (Table S5). Amino acid frequency analyses revealed that the highest frequencies were leucine and isoleucine, whereas cysteine was a rare amino acid. The codon exhibited a strong bias toward an A or T at the third position. An RSCU value below 1.00 indicates that the codon usage frequency is lower than expected, whereas an RSCU value above 1.00 indicates that the codon usage frequency is higher than expected. In this study, the RSCU values of 30 codons were greater than 1, whereas the RSCU value of 32 codons was less than 1, and 2 codons were equal to 1 (Figure 2).

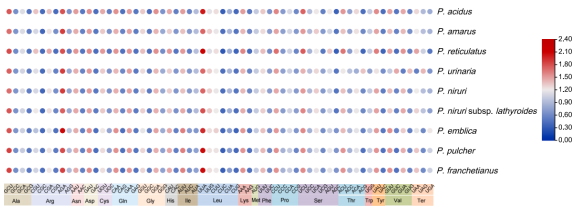


FIGURE 2
The RSCU values of nine *Phyllanthus* cp genomes. Color key: the red values indicate higher RSCU values, and the blue values indicate lower RSCU values (For interpretation of the references to color in this Figure legend, the reader is referred to the web version of this article).

Moreover, the results showed that the GC content of synonymous third codon positions (GC3s) is closely related to codon bias, and the values of GC3s ranged from 25.0% to 31.1%, suggesting that the genus *Phyllanthus* had a greater preference for the A/U ending codons. And the GC content of these cp genomes was highly conserved. In addition, the values for the effective number of codons ranged from 48.43% to 52.93%. Both the codon adaptation index and optimal frequency were less than 0.5. These findings indicated a slight bias toward codon usage in the nine *Phyllanthus* species.

3.4 Inverted repeats

Expansion and contraction at the borders of IR regions are common evolutionary phenomena that may explain variations in the size of cp genomes. As illustrated in Figure 3, the *rps3* genes existed entirely in the LSC regions of all species, and *rpl2* existed entirely in the IR region except for *P. niruri* subsp. *lathyroides*. A truncated copy of the *rpl22* gene was observed at the junction of LSC/IRb in three species (*P. reticulatus*, *P. pulcher*, and *P. franchetianus*), which starts in LSC regions and integrates into the IRb region with a size ranging from 2 to 23 bp, whereas the remaining six species were present entire in the LSC region. Another truncated copy of the *rps19* gene was found at the junction of IRa/SSC in two species (*P. acidus* and *P. emblica*). Notably, *rps19* exists entirely in the LSC region for *P. niruri* subsp. *lathyroides*, and *rps19* is present entirely in the IR region in the remaining six species. Besides, the *ndhF* gene was found in the SSC regions except for three species (*P. acidus*, *P. pulcher*, and *P. franchetianus*), which start in the SSC regions and integrate into the IRb region in those species. Moreover, the *ycf1* gene was observed at the IRa/SSC junction except for *P. urinaria*. In all other species, the *ycf1* gene starts in SSC regions and integrates into the IRa. However, in *P. urinaria*, the *ycf1* gene is completely present in the SSC region. Both *psbA* and *trnN* exist entirely in the LSC and IRa, respectively. Notably, the *trnL* gene exists only in the IR region of *P. niruri* subsp. *lathyroides*. These results show that the cp genomes of nine *Phyllanthus* species display a unique IR contraction and expansion type.

3.5 Genome comparison and nucleotide diversity

A comparison of overall sequence variation showed that the cp genome of *Phyllanthus* is quite different. The sequence divergence of IR regions was lower than that of SSC and LSC regions, and the coding region was more conserved than the non-coding regions. Except for the more remarkable mutations in the *ndhF*, *ycf1*, and *ycf2* genes, all protein-coding genes showed a highly conserved character. The highest divergence was mainly found in intergenic

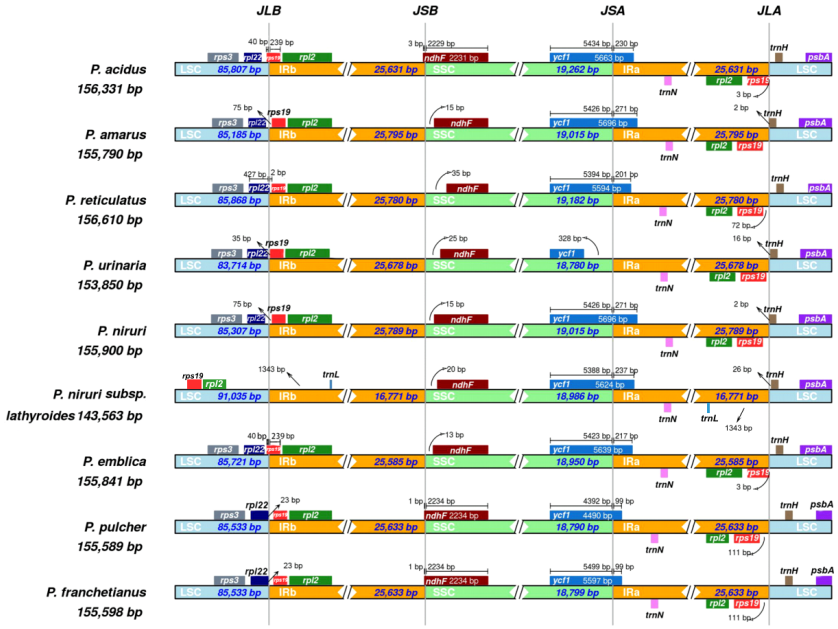


FIGURE 3
Comparisons of the borders of LSC, SSC, and IRa/b regions among the nine *Phyllanthus* cp genomes. The numbers represent the distance between the gene ends and the border sites, and the numbers below represent the length of the LSC, SSC, and IRa/b regions.

spacers (IGS), such as *rps16-trnQ-UUG*, *trnS-GCU-trnG-UCC*, *trnG-UCC-trnR-UCU*, *trnE-UUC-trnT-GGU*, *trnD-GUC-trnY-GUA*, *trnT-UGU-trnL-UAA*, *trnL-CAA-ndhB*, *trnN-GUU-trnR-ACG*, and *rps15-ycf1* (Figure 4). Besides, the sliding window analysis demonstrated that the nine regions, including *rps16*, *trnS-GCU-trnG-UCC*, *trnG-UCC-trnR-UCU*, *petA-psbJ*, *rps3*, *rrn5S-trnR-ACG*, *ndhF*, *ndhE-ndhG*, and *ycf1*, had higher nucleotide diversity values (Pi > 0.05) (Figure S3). The results above show that 12 highly variable sites (*rps16-trnQ-UUG*, *trnS-GCU-trnG-UCC*, *trnG-UCC-trnR-UCU*, *trnE-UUC-trnT-GGU*, *trnD-GUC-trnY-GUA*, *trnT-UGU-trnL-UAA*, *trnL-CAA-ndhB*, *trnN-GUU-trnR-ACG*, *rps15-ycf1*, *petA-psbJ*, *rrn5S-trnR-ACG*, and *ndhE-ndhG*) might be able to be used as molecular markers to identify *P. urinaria* and its contaminants.

3.6 Species authentication analysis based on IGS

IGS regions are the most commonly used markers for phylogenetic studies at plant taxonomic levels, as they are regarded as more variable and may provide more phylogenetically informative characters. To find candidate sequences for identifying *P. urinaria* and its adulterants, 12 IGS were extracted from 13 *Phyllanthus* species using PhyloSuite v1.2.2. And each of them is subject to ML analyses in IQtree. As shown in Figure S4.1-4.12, *P. urinaria* could be distinguished from its common adulterants based on *trnS-GCU-trnG-UCC*, *trnT-UGU-trnL-UAA*, and *petA-psbJ*, whereas the remaining IGS cannot be distinguished, and the bootstrap values for the relationship among these clades were weak (<70%). Furthermore, the ML phylogenetic tree was also inferred using a combination of these three IGS. The results (Figure S5) showed that *P. urinaria* (Genbank OP009346) was located in independent branches and that there was a well-supported sister relationship between *P. urinaria* (Genbank OP009346) and *P. amarus* (Genbank OP009344) + *P. urinaria* (Genbank NC060522) + *P. niruri* (Genbank OP009347). These results indicated that combining three IGS could effectively discriminate *P. urinaria* from its common adulterants.

3.7 Phylogenetic analysis and divergence time analysis

The ML and NJ phylogenetic trees were inferred using 40 species, with two *Daphniphyllum* species as outgroups. The consensus trees obtained from the inference analyses showed that most nodes resolved with high support (Figure 5, Figure S6). The phylogenetic trees generated by the ML and NJ alignments have similar topologies. Each subfamily of the Euphorbiaceae family forms a monophyletic clade. Acalyphoideae and Euphorbioideae + Crotonoideae were sister taxa within the four subfamilies, and Phyllanthoideae was a sister group to the clade containing Acalyphoideae + Crotonoideae +

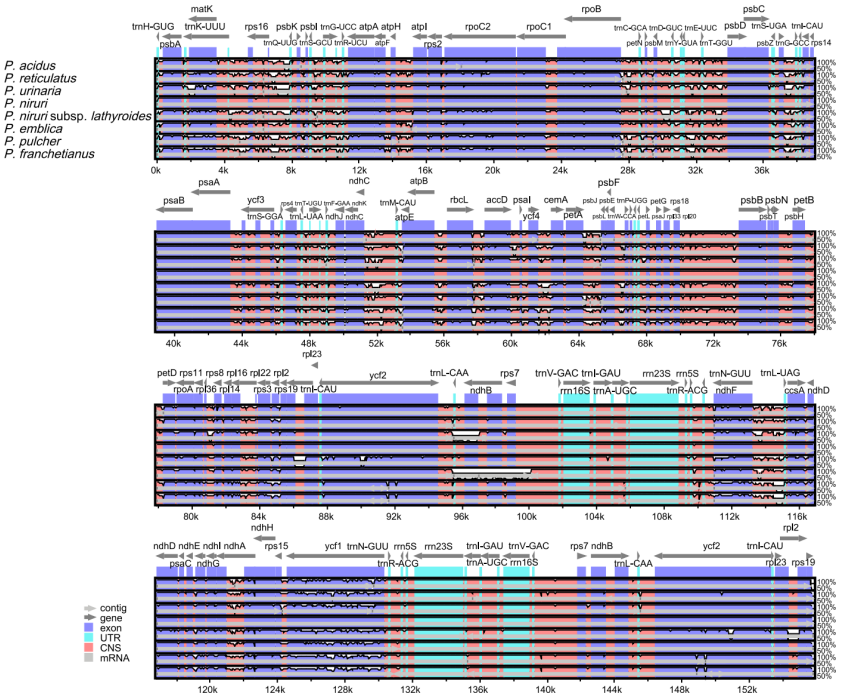


FIGURE 4
Comparison of nine cp genomes using *P. amarus* annotation as a reference. The vertical scale indicates the percentage of identity, ranging from 50% to 100%. The horizontal axis shows the coordinates within the cp genome. Genome regions are color-coded as exons, introns, and IGS, and the gray arrows indicate the direction of transcription of each gene. Annotated genes are displayed along the top.

Euphorbioideae. In addition, the subfamily Phyllanthoideae is further divided into three clades: (i) clade A, including *Glochidion wrightii* (Genbank MW801302), *G. eriocarpum* (Genbank MW801303), *G. chodoense* (Genbank NC042906), *G. puberum* (Genbank MW801304), *P. amarus* (Genbank NC047474), *P. emblica* (Genbank MN122078, Genbank OP0 09349, Genbank NC047477), *P. acidus* (Genbank OP009343), and *P. niruri* subsp. *lathyroides* (Genbank OP009348); (ii) clade B, including *P. amarus* (Genbank OP009344), *P. urinaria* (Genbank NC060522), *P. niruri* (Genbank OP009347), and *P. urinaria* (Genbank OP009346); (iii) clade C, including *P. franchetianus* (Genbank OP009351), *P. pulcher* (Genbank OP009350), and *P. reticulatus* (Genbank OP009345).

In addition, 34 cp genomes of Euphorbiaceae family plants (including 13 *Phyllanthus* species) and two outgroups are used to estimate the divergence times. *Phyllanthus* were estimated to have originated 48.72 million years ago (Ma). The two main lineages, clade A + clade B and clade C, seem to have radiated since the Oligocene (clade A + clade B: 36.85 Ma; clade C: 30.89 Ma; Figure 6). The extant subfamilies of the Euphorbioideae and Acalyphoideae shared a common ancestor at the beginning of the Cretaceous (91.91 Ma), while the split between the Euphorbioideae and Crotonoideae is estimated to have occurred at 75.64 Ma. Moreover, the divergence times of the *Daphniphyllum* occurred at 120.77 Ma, having shared a common ancestor with the Euphorbiaceae.

4 Discussion

4.1 Cp genome structure and comparative analysis

In the present study, the cp genomes of nine species of *Phyllanthus* were assembled *de novo* and analyzed comparatively. The observed genome size is within the size range known for most angiosperms, ranging from 107 kb in *Cathaya argyrophylla* (Pinaceae) to 218 kb in *Pelargonium* (geraniums; Geraniaceae) (Daniell et al., 2016). Besides, our findings revealed that the total length, GC content, and gene composition of the cp genomes were almost identical in all species. Previous studies have shown that the cp genome of angiosperms is highly conserved at the genus level (Khan et al., 2020; Villanueva-Corrales et al., 2021; Feng et al., 2022). Moreover, previous research demonstrated that the *rps19* gene existed in the IR region (Ahmed et al., 2012; Wang et al., 2022). In contrast, the *rps19* of *P. niruri* subsp. *lathyroides* is located entirely in the LSC Region, possibly due to IR contraction. The same pattern was reported in *Anchomanes hookeri*, and *Peucedanum* (Henriquez et al., 2020; Liu et al., 2022).

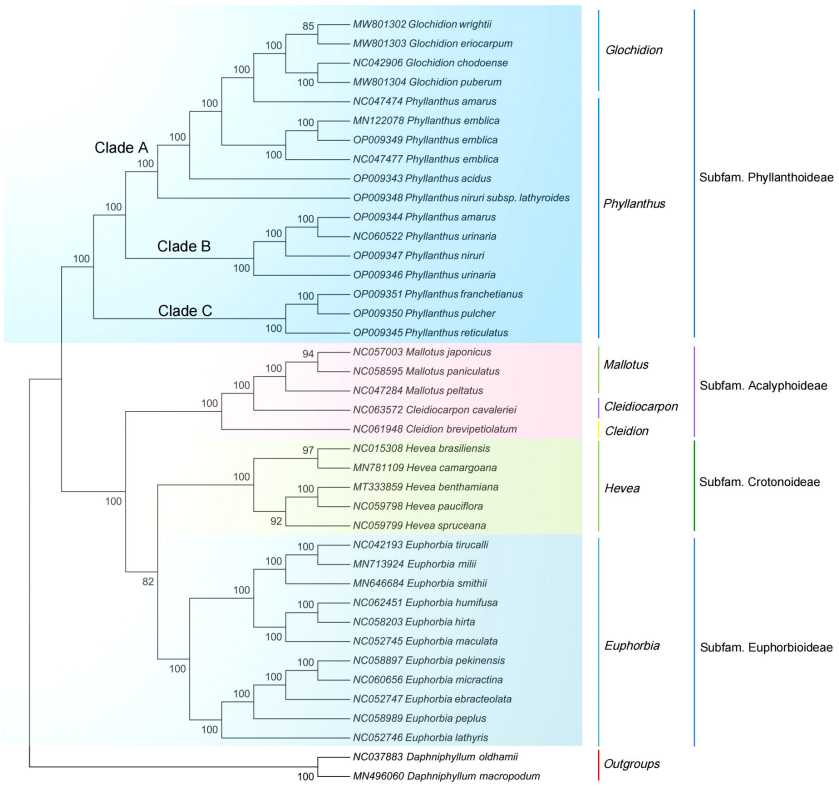


FIGURE 5
Maximum likelihood phylogenetic tree based on complete cp genomes. *Daphniphyllum oldhamii* and *D. macropodum* were used as outgroups. Numbers at nodes are bootstrap support values.

4.2 Species identification and phylogenetic analysis

Previous molecular studies of the *Rheum*, *Hedyotis*, and *Curcuma* species showed that cp genetic markers had high identification capabilities (Zhou et al., 2018; Gui et al., 2020; Yik et al., 2021). Our study revealed that three regions (*trnS-GCU-trnG-UCC*, *trnT-UGU-trnL-UAA*, and *petA-psbJ*) might be potential molecular markers for identifying *P. urinaria* and its common adulterants. Bouman et al. (2021) found that the *trnS-GCU-trnG-UCC* could distinguish *Phyllanthus* species. Notably, Zhang et al. (2021) also found that the *trnS-GCU-trnG-UCC* could be potential molecular markers for distinguishing *Alpinia* species. Moreover, *trnT-UGU-trnL-UAA* or *petA-psbJ* were reported as potential markers for other species identification (Dong et al., 2021; Wu et al., 2021). Although these previous studies revealed that universal DNA barcode (e.g., *psbA-trnH*) could differentiate *P. urinaria* from their related species (Srirama et al., 2010; Inglis et al., 2018), some common adulterants were not included in these studies. Furthermore, the comparative analysis showed that the screened IGS exhibited higher variability than *psbA-trnH*. Theoretically, these IGS could differentiate nine selected species, whereas a much more thorough investigation of identification accuracy and amplification efficiency is required, as well as more experimental evidence.

Moreover, ML analysis demonstrated that *P. urinaria* (Genbank OP009346) was located in independent branches in the phylogeny and strongly supported the sister relationship between *P. urinaria* (Genbank OP009346) and the well-supported clade (*P. amarus*, Genbank OP009344; and *P. niruri*, Genbank OP009347). The results indicated that the cp genome could discriminate *P. urinaria* from *P. amarus* and *P. niruri*, which was supported by the findings of other researchers based on ITS, *matK*, *psbA-trnH*, *trnL*, and *trnL-trnF* (Inglis et al., 2018). However, the samples of *P. urinaria* (Genbank OP009346) and *P. urinaria* (Genbank NC060522) were not recovered as monophyletic and were placed in different branches. In the previous study, some researchers found that intraspecific diversity existed in *Isodon rubescens* and *Artemisia argyi* from different geographical areas (Zhou et al., 2022; Chen et al., 2022). Therefore, the difference in geographical origins may explain why the two species are split in these clades. Besides, both NJ and ML analyses found strong support for a sister relationship between *P. reticulatus* (Genbank OP009345) and *P. pulcher* (Genbank OP009350), which agreed with the findings of Hidalgo et al. (2020) based on *trnK-matK*, *matK*, ITS, and *matK*+ITS

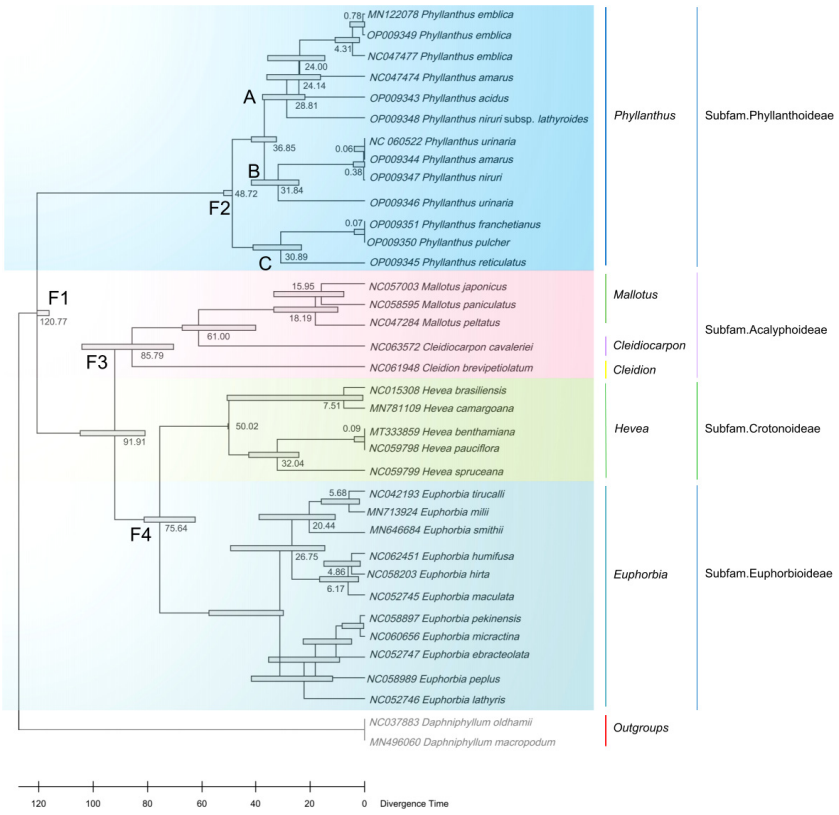


FIGURE 6
Divergence times estimation based on cp genomes. The node ages are given for each node.

(Pagare et al., 2016). In a previous study, Rehman et al. (2021) also found that the polymorphic protein-coding genes, including *rpl22*, *ycf1*, *matK*, *ndhF*, and *rps15*, may help reconstruct the high-resolution phylogenetic tree of the family Phyllanthaceae. In general, our results provide a valuable reference and a foundation for using cp genomes in species identification and aid in the understanding of the phylogeny of *Phyllanthus*.

4.3 Divergence time of *Phyllanthus*

According to divergence time estimates, the early divergence of *Phyllanthus* occurred at approximately 48.72 Ma during the early Eocene, which is congruent with other studies (Kawakita and Kato, 2009; Welzen et al., 2015). Since the late Eocene, the previous study reported that the global climate started to have a notable change; as the humidity and precipitation gradually increased (Zachos et al., 2001) and slowly cooled within this timeframe. These climate changes may have promoted the dispersals/migrations, and diversification of land plants (Zuo et al., 2017). In addition, Dynesius and Jansson (2000) and Zachos et al. (2001) also found that the temperature increase affected various plant and animal communities and groups at the Oligocene/Miocene boundary (~23 Ma). Among the effects of climate change, there was likely increased speciation in the niches that opened after the end of the climatic fluctuations (Bouman et al., 2021). Therefore, we can conclude that the climatic changes may have contributed to the diversification of *Phyllanthus* during the Eocene.

5 Conclusion

In the present study, the complete cp genomes of nine species of *Phyllanthus* were *de novo* assembled from high throughput sequencing reads, and the cp genomes of *P. acidus*, *P. reticulatus*, *P. niruri*, *P. niruri* subsp. *lathyroides*, *P. pulcher*, and *P. franchetianus* were reported for the first time. These cp genomes were generally conserved and exhibited similar gene content and genomic structure. Three highly variable cp loci, including *trnS-GCU-trnG-UCC*, *trnT-UGU-trnL-UAA*, and *petA-psbJ* were identified and could serve as candidate markers for identifying *P. urinaria* and its common adulterants. Meanwhile, the complete cp genome was considered a reliable molecular marker for identifying these species, which may have virtual significance for protecting their diversity and making management decisions

for this species. The divergence of *Phyllanthus* from ancestral taxa occurred in the early Eocene, which might be due to geological and climatic changes. In conclusion, our study provides a powerful tool and valuable scientific reference for the safety and effectiveness of clinical drug use, and it also contributes to the bioprospecting and conservation of *Phyllanthus* species.

Data availability statement

The data presented in the study are deposited in the GenBank repository, accession numbers were from OP009343 to OP009351.

Author contributions

HF, YL, and PZ participated in the conception and design of the research. HF, GD and BL collected the species. HF and GD are responsible for analyzing and processing data. HF wrote the manuscript. The paper was revised by YL and PZ. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Yunnan academician expert workstation (202105AF150053), the key technology projects in the Yunnan province of China (202002AA100007).

Acknowledgments

We would like to thank Mr. Haitao Li for his assistance in obtaining specimens for this study. We also thank Yuan Jiang helped to teach the software used for the experiments.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1099856/full#supplementary-material

References

Abdullah,, Mehmood, F., Shahzadi, I., Waseem, S., Mirza, B., Ahmed, I., et al. (2020). Chloroplast genome of hibiscus rosa-sinensis (Malvaceae): Comparative analyses and identification of mutational hotspots. *Genomics* 112, 581–591. doi: 10.1016/j.ygeno.2019.04.010

Adedapo, A. A., Abatan, M. O., Idowu, S. O., and Olorunsogo, O. O. (2005). Toxic effects of chromatographic fractions of *Phyllanthus amarus* on the serum biochemistry of rats. *Phytother. Res.* 19, 812–815. doi: 10.1002/ptr.1721

Ahmed, I., Biggs, P. J., Matthews, P. J., Collins, L. J., Hendy, M. D., and Lockhart, P. J. (2012). Mutational dynamics of aroid chloroplast genomes. *Genome Biol. Evol.* 4, 1316–1323. doi: 10.1093/gbe/evs110

Allen, G. C., Flores-Vergara, M. A., Krasynanski, S., Kumar, S., and Thompson, W. F. (2006). A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. *Nat. Protoc.* 1, 2320–2325. doi: 10.1038/nprot.2006.384

Amiryousefi, A., Hyvönen, J., and Poczai, P. (2018). IRscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics* 34, 3030–3031. doi: 10.1093/bioinformatics/bty220

Beier, S., Thiel, T., Münch, T., Scholz, U., and Mascher, M. (2017). MISA-web: A web server for microsatellite prediction. *Bioinformatics* 33, 2583–2585. doi: 10.1093/bioinformatics/btx198

Bouman, R. W., Kessler, P. J. A., Telford, I. R. H., Bruhl, J. J., Strijk, J. S., Saunders, R. M. K., et al. (2021). Molecular phylogenetics of *Phyllanthus* sensu lato (Phyllanthaceae): towards coherent monophyletic taxa. *Taxon* 70, 72–98. doi: 10.1002/tax.12424

Brudno, M., Malde, S., Poliakov, A., Do, C. B., Couronne, O., Dubchak, I., et al. (2003). Glocal alignment: Finding rearrangements during alignment. *Bioinformatics* 19, i54–i62. doi: 10.1093/bioinformatics/btg1005

Chase, M. W., Christenhusz, M. J. M., Fay, M. F., Byng, J. W., Judd, W. S., Soltis, D. E., et al. (2016). An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG IV. *Bot. J. Linn. Soc* 181, 1–20. doi: 10.1111/boj.12385

Chen, C. J., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y. H., et al. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp.2020.06.009

Chen, C. J., Miao, Y. H., Luo, D. D., Li, J. X., Wang, Z. X., Luo, M., et al. (2022). Sequence characteristics and phylogenetic analysis of the *Artemisia argyi* chloroplast genome. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.906725

Chudapongse, N., Kamkhunthod, M., and Poompachee, K. (2010). Effects of *Phyllanthus urinaria* extract on HepG2 cell viability and oxidative phosphorylation by isolated rat liver mitochondria. *J. Ethnopharmacol.* 130, 315–319. doi: 10.1016/j.jep.2010.05.010

Daniell, H., Lin, C. S., Yu, M., and Chang, W. J. (2016). Chloroplast genomes: Diversity, evolution, and applications in genetic engineering. *Genome Bio.* 17, 134. doi: 10.1186/s13059-016-1004-2

Dong, W. P., Liu, Y. L., Xu, C., Gao, Y. W., Yuan, Q. J., Suo, Z. L., et al. (2021). Chloroplast phylogenomic insights into the evolution of *Distylium* (Hamamelidaceae). *BMC Genomics* 22, 293. doi: 10.1186/s12864-021-07590-6

Dynesius, M., and Jansson, R. (2000). Evolutionary consequences of changes in species' geographical distributions driven by milankovitch climate oscillations. *P. Natl. Acad. Sci. U.S.A.* 97, 9115–9120. doi: 10.1073/pnas.97.16.9115

Feng, Y., Gao, X. F., Zhang, J. Y., Jiang, L. S., Li, X., Deng, H. N., et al. (2022). Complete chloroplast genomes provide insights into evolution and phylogeny of *Campylotropis* (Fabaceae). *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.895543

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., and Dubchak, I. (2004). VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* 32, W273–W279. doi: 10.1093/nar/gkh458

Geethangili, M., and Ding, S. (2018). A review of the phytochemistry and pharmacology of *Phyllanthus urinaria l. Front. Pharmacol.* 9. doi: 10.3389/fphar.2018.01109

Geng, H. C., Zhu, H. T., Yang, W. N., Wang, D., Yang, C. R., and Zhang, Y. J. (2021). New cytotoxic dichapetalins in the leaves of *Phyllanthus acidus*: Identification, quantitative analysis, and preliminary toxicity assessment. *Bioorg. Chem.* 114, 105125. doi: 10.1016/j.bioorg.2021.105125

Greiner, S., Lehwark, P., and Bock, R. (2019). OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* 47, W59–W64. doi: 10.1093/nar/gkz238

Guan, Y. H., Liu, W. W., Duan, B. Z., Zhang, H. Z., Chen, X. B., Wang, Y., et al. (2022). The first complete chloroplast genome of Vicatia thibetica de boiss.: Genome features, comparative analysis, and phylogenetic relationships. *Physiol. Mol. Biol. Plants* 28, 439–454. doi: 10.1007/s12298-022-01154-y

Gui, L. J., Jiang, S. F., Xie, D. F., Yu, L. Y., Huang, Y., Zhang, Z. J., et al. (2020). Analysis of complete chloroplast genomes of *Curcuma* and the contribution to phylogeny and adaptive evolution. *Gene* 732, 144355. doi: 10.1016/j.gene.2020.144355

Guo, Q., Zhang, Q. Q., Chen, J. Q., Zhang, W., Qiu, H. C., Zhang, Z. J., et al. (2017). Liver metabolomics study reveals protective function of *Phyllanthus urinaria* against CCl4-induced liver injury. *Chin. J. Nat. Medicines* 15, 525–533. doi: 10.1016/S1875-5364(17)30078-X

Heinze, B. (2007). A database of PCR primers for the chloroplast genomes of higher plants. *Plant Methods* 3, 4. doi: 10.1186/1746-4811-3-4

Henriquez, C. L., Abdullah,, Ahmed, I., Carlsen, M. M., Zuluaga, A., Croat, T. B., et al. (2020). Evolutionary dynamics of chloroplast genomes in subfamily *Aroideae* (Araceae). *Genomics* 112, 2349–2360. doi: 10.1016/j.ygeno.2020.01.006

Hidalgo, B. F., Bazan, S. F., Iturralde, R. B., and Borsch, T. (2020). Phylogenetic relationships and character evolution in neotropical *Phyllanthus* (Phyllanthaceae), with a focus on the Cuban and Caribbean taxa. *Int. J. Plant Sci.* 181, 284–305. doi: 10.1086/706454

Inglis, P., Mata, L., Da Silva, M., Vieira, R., Alves, R. D. B. N., Silva, D., et al. (2018). DNA Barcoding for the identification of *Phyllanthus* taxa used medicinally in Brazil. *Planta Med.* 84, 1300–1310. doi: 10.1055/a-0644-2688

Jiang, Y., Miao, Y. J., Qian, J., Zheng, Y., Xia, C. L., Yang, Q. S., et al. (2022). Comparative analysis of complete chloroplast genome sequences of five endangered species and new insights into phylogenetic relationships of *Paris*. *Gene* 833, 146572. doi: 10.1016/j.gene.2022.146572

Jin, J. J., Yu, W. B., Yang, J. B., Song, Y., DePamphilis, C. W., Yi, T. S., et al. (2020). GetOrganelle: a fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. *Genome Biol.* 21, 241–272. doi: 10.1186/s13059-020-02154-5

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth.4285

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kawakita, A., and Kato, M. (2009). Repeated independent evolution of obligate pollination mutualism in the phyllantheae-epicephala association. *P Roy Soc. B-Biol Sci.* 276, 417–426. doi: 10.1098/rspb.2008.1226

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649. doi: 10.1093/bioinformatics/bts199

Khan, A., Asaf, S., Khan, A. L., Shehzad, T., Al-Rawahi, A., and Al-Harrasi, A. (2020). Comparative chloroplast genomics of endangered euphorbia species: Insights into hotspot divergence, repetitive sequence variation, and phylogeny. *Plants* 9, 199. doi: 10.3390/plants9020199

Kiran, K. R., Swathy, P. S., Paul, B., Shama Prasada, K., Radhakrishna Rao, M., Joshi, M. B., et al. (2021). Untargeted metabolomics and DNA barcoding for discrimination of *Phyllanthus* species. *J. Ethnopharmacol.* 273, 113928. doi: 10.1016/j.jep.2021.113928

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096

Kumar, S., Stecher, G., Suleski, M., and Hedges, S. B. (2017). TimeTree: A resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* 34, 1812–1819. doi: 10.1093/molbev/msx116

Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* 29, 4633–4642. doi: 10.1093/nar/29.22.4633

Li, F., Liu, Y., Wang, J. H., Xin, P. Y., Zhang, J. Y., Zhao, K., et al. (2022). Comparative analysis of chloroplast genome structure and phylogenetic relationships among six taxa within the genus *Catalpa* (Bignoniaceae). *Front. Genet.* 13. doi: 10.3389/fgene.2022.845619

Li, X. W., Yang, Y., Henry, R. J., Rossetto, M., Wang, Y. T., and Chen, S. L. (2015). Plant DNA barcoding: From gene to genome. *Biol. Rev.* 90, 157–166. doi: 10.1111/brv.12104

Liu, C. K., Lei, J. Q., Jiang, Q. P., Zhou, S. D., and He, X. J. (2022). The complete plastomes of seven *Peucedanum* plants: comparative and phylogenetic analyses for the peucedanum genus. *BMC Plant Biol.* 22, 101. doi: 10.1186/s12870-022-03488-x

Manissorn, J., Sukrong, S., Ruangrungsi, N., and Mizukami, H. (2010). Molecular phylogenetic analysis of *Phyllanthus* species in Thailand and the application of polymerase chain reaction-restriction fragment length polymorphism for *Phyllanthus* amarus identification. *Biol. Pharm. Bull.* 33, 1723–1727. doi: 10.1248/bpb.33.1723

Mello, B. (2018). Estimating TimeTrees with MEGA and the TimeTree resource. *Mol. Biol. Evol.* 35, 2334–2342. doi: 10.1093/molbev/msy133

Neuhaus, H. E., and Emes, M. J. (2000). Nonphotosynthetic metabolism in plastids. *Annu. Rev. Plant Biol.* 51, 111–140. doi: 10.1146/annurev.arplant.51.1.111

Nguyen, L., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2015). IQ-Tree: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Pagare, R., Naik, S., Krishnan, S., and Janarthanam, M. (2016). Lectotypification of pseudoglochidion anamalayanum gamble and its taxonomic position under the genus *Phyllanthus* (Phyllanthaceae). *Phytotaxa* 286, 61–74. doi: 10.11646/phytotaxa.286.2.1

Pruesapan, K., Telford, I. R. H., Bruhl, J. J., and van Welzen, P. C. (2012). Phylogeny and proposed circumscription of *Breynia*, *Sauropus* and *Synostemon* (Phyllanthaceae), based on chloroplast and nuclear DNA sequences. *Aust. Syst. Bot.* 25, 313. doi: 10.1071/SB11005

Rehman, U., Sultana, N., Abdullah,, Jamal, A., Muzaffar, M., and Poczai, P. (2021). Comparative chloroplast genomics in *Phyllanthaceae* species. *Diversity (Basel)* 13, 403. doi: 10.3390/d13090403

Ren, J., Tian, J., Jiang, H., Zhu, X. X., Mutie, F. M., Wanga, V. O., et al. (2022). Comparative and phylogenetic analysis based on the chloroplast genome of *Coleanthus subtilis* (Tratt.) Seidel, a protected rare species of monotypic genus. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.828467

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., et al. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol. Biol. Evol.* 34, 3299–3302. doi: 10.1093/molbev/msx248

Sarin, B., Clemente, J. P. M., and Mohanty, A. (2013). PCR-RFLP to distinguish three *Phyllanthus* sp., commonly used in herbal medicines. *S. Afr. J. Bot.* 88, 455–458. doi: 10.1016/j.sajb.2013.09.011

Schattner, P., Brooks, A. N., and Lowe, T. M. (2005). The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* 33, W686–W689. doi: 10.1093/nar/gki366

Sharp, P. M., Tuohy, T. M., and Mosurski, K. R. (1986). Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res.* 14, 5125–5143. doi: 10.1093/nar/14.13.5125

Shi, L. C., Chen, H. M., Jiang, M., Wang, L. Q., Wu, X., Huang, L. F., et al. (2019). CPGAVAS2, an integrated plastome sequence annotator and analyzer. *Nucleic Acids Res.* 47, W65–W73. doi: 10.1093/nar/gkz345

Srirama, R., Senthilkumar, U., Sreejayan, N., Ravikanth, G., Gurumurthy, B. R., Shivanna, M. B., et al. (2010). Assessing species admixtures in raw drug trade of *Phyllanthus*, a hepato-protective plant using molecular tools. *J. Ethnopharmacol.* 130, 208–215. doi: 10.1016/j.jep.2010.04.042

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., et al. (2017). GeSeq-versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* 45, W6–W11. doi: 10.1093/nar/gkx391

Villanueva-Corrales, S., García-Botero, C., Garcés-Cardona, F., Ramírez-Ríos, V., Villanueva-Mejía, D. F., and Álvarez, J. C. (2021). The complete chloroplast genome of *Plukenetia volubilis* provides insights into the organelle inheritance. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.667060

Wang, J., Qian, J., Jiang, Y., Chen, X. C., Zheng, B. J., Chen, S. L., et al. (2022). Comparative analysis of chloroplast genome and new insights into phylogenetic relationships of *Polygonatum* and tribe *Polygonateae*. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.882189

Webster, G. L., and Carpenter, K. J. (2008). Pollen morphology and systematics of palaeotropical *Phyllanthus* and related genera of subtribe *Phyllanthinae* (Euphorbiaceae). *Bot. J. Linn. Soc* 157, 591–608. doi: 10.1111/j.1095-8339.2008.00781.x

Welzen, P. C., Pruesapan, K., Telford, I. R. H., and Bruhl, J. J. (2015). Historical biogeography of *Breynia* (Phyllanthaceae); what caused speciation? *J. Biogeogr.* 42, 1493–1502. doi: 10.1111/jbi.12517

Wu, L. W., Cui, Y. X., Wang, Q., Xu, Z. C., Wang, Y., Lin, Y. L., et al. (2021). Identification and phylogenetic analysis of five *Crataegus* species (Rosaceae) based on complete chloroplast genomes. *Planta* 254, 14. doi: 10.1007/s00425-021-03667-4

Xiong, C., Sun, W., Li, J. J., Yao, H., Shi, Y. H., Wang, P., et al. (2018). Identifying the species of seeds in traditional Chinese medicine using DNA barcoding. *Front. Pharmacol.* 9. doi: 10.3389/fphar.2018.00701

Yik, M. H., Kong, B. L., Siu, T., Lau, D. T., Cao, H., and Shaw, P. (2021). Differentiation of *Hedyotis diffusa* and common adulterants based on chloroplast genome sequencing and DNA barcoding markers. *Plants* 10, 161. doi: 10.3390/plants10010161

Zachos, J., Pagani, M., Sloan, L., Thomas, E., and Billups, K. (2001). Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* 292, 686–693. doi: 10.1126/science.1059412

Zhang, Y., Song, M. F., Li, Y., Sun, H. F., Tang, D. Y., Xu, A. S., et al. (2021). Complete chloroplast genome analysis of two important medicinal *Alpinia* species: *Alpinia galanga* and *Alpinia kwangsiensis*. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.705892

Zhou, Y. X., Nie, J., Xiao, L., Hu, Z. G., and Wang, B. (2018). Comparative chloroplast genome analysis of rhubarb botanical origins and the development of specific identification markers. *Molecules* 23, 2811. doi: 10.3390/molecules 23112811

Zhou, Z. Y., Wang, J., Pu, T. T., Dong, J. J., Guan, Q., Qian, J., et al. (2022). Comparative analysis of medicinal plant *Isodon rubescens* and its common adulterants based on chloroplast genome sequencing. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.1036277

Zuo, Y. J., Wen, J., and Zhou, S. L. (2017). Intercontinental and intracontinental biogeography of the eastern Asian-Eastern north American disjunct *Panax* (the ginseng genus, *Araliaceae*), emphasizing its diversification processes in eastern Asia. *Mol. Phylogenet. Evol.* 117, 60–74. doi: 10.1016/j.ympev.2017.06.016

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 13 February 2023
DOI 10.3389/fpls.2023.1119041

OPEN ACCESS

EDITED BY
Weijun Kong,
Capital Medical University, China

REVIEWED BY
Zhichao Xu,
Northeast Forestry University, China
Huasheng Peng,
National Resource Center for Chinese Materia Medica, China Academy of Chinese Medical Sciences, China
Wei Sun,
Key Laboratory of Beijing for Identification and Safety Evaluation of Chinese Medicine, China Academy of Chinese Medical Sciences, China

*CORRESPONDENCE
Jinping Han
happymyra2007@163.com

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

RECEIVED 08 December 2022
ACCEPTED 20 January 2023
PUBLISHED 13 February 2023

CITATION
Bai X, Wang G, Ren Y, Su Y and Han J (2023) Insights into taxonomy and phylogenetic relationships of eleven *Aristolochia* species based on chloroplast genome.
Front. Plant Sci. 14:1119041.
doi: 10.3389/fpls.2023.1119041

Insights into taxonomy and phylogenetic relationships of eleven *Aristolochia* species based on chloroplast genome

Xuanjiao Bai†, Gang Wang†, Ying Ren, Yuying Su and Jinping Han*

Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

Introduction: The Aristolochia, as an important genus comprised of over 400 species, has attracted much interest because of its unique chemical and pharmacological properties. However, the intrageneric taxonomy and species identification within *Aristolochia* have long been difficult because of the complexity of their morphological variations and lack of high-resolution molecular markers.

Methods: In this study, we sampled 11 species of *Aristolochia* collected from distinct habitats in China, and sequenced their complete chloroplast (cp) genomes.

Results: The 11 cp genomes of *Aristolochia* ranged in size from 159,375bp (*A. tagala*) to 160,626 bp (*A. tubiflora*), each containing a large single-copy (LSC) region (88,914-90,251 bp), a small single-copy (SSC) region (19,311-19,917 bp), and a pair of inverted repeats (IR) (25,175-25,698 bp). These cp genomes contained 130-131 genes each, including 85 protein-coding genes (CDS), 8 ribosomal RNA genes, and 37-38 transfer RNA genes. In addition, the four types of repeats (forward, palindromic, reverse, and complement repeats) were examined in *Aristolochia* species. *A. littoralis* had the highest number of repeats (168), while *A. tagala* had the lowest number (42). The total number of simple sequence repeats (SSRs) is at least 99 in *A. kwangsiensis*, and, at most, 161 in *A. gigantea*. Interestingly, we detected eleven highly mutational hotspot regions, including six gene regions (*clpP, matK, ndhF, psbT, rps16, trnK*-UUU) and five intergenic spacer regions (*ccsA-ndhD, psbZ-trnG*-GCC, *rpl33-rps18, rps16-trnQ*-UUG, *trnS*-GCU-*trnG*-UCC). The phylogenetic analysis based on the 72 protein-coding genes showed that 11 *Aristolochia* species were divided into two clades which strongly supported the generic segregates of the subgenus *Aristolochia* and *Siphisia*.

Discussion: This research will provide the basis for the classification, identification, and phylogeny of medicinal plants of Aristolochiaceae.

KEYWORDS

Aristolochia, taxonomy, phylogenetic relationship, comparative analysis, chloroplast genome

1 Introduction

Aristolochia, a type genus of the family Aristolochiaceae, is widely distributed in tropical, subtropical, and temperate areas. Approximately 45 species are distributed in China, and 33 are endemic (Huang et al., 2003). Many species of *Aristolochia* possess a long history of medicinal value. For example, *A. manshuriensis* was commonly used as a traditional Chinese medicine to alleviate pathogenic fire. The dry mature fruits of *A. contorta* and *A. debilis* were called "Fructus Aristolochiae" and had been used to relieve cough and alleviate hemorrhoids. Else species such as *A. fangchi*, *A. tagala*, and *A. kwangsiensis* are widely used in folk medicine and are important medicinal plants. However, the outbreak of renal disease among the group of young women who followed the same slimming medicine containing *A. fangchi* sounds an alarm about the delayed toxic effects of *Aristolochia* species (Vanherweghem et al., 1993; Tomlinson et al., 2020). After decades of investigation, increasing research verified the aristolochic acid contained in the *Aristolochia* species was the main causative factor of nephropathy and may be the potential to cause cancer (Stefanovic et al., 2006; Jelaković et al., 2019). Hence, the *Aristolochia* species have been excluded from the Chinese pharmacopeia and banned to utilize for medicinal purposes in many countries (Kim and Lim, 2019). Yet the conflict between the medicinal value and potential nephrotoxicity and teratogenicity makes the illegal addition of *Aristolochia* in medicines and health products still rampant (Maggini et al., 2018; Ji et al., 2021). Recently, modern studies gradually discovered the new bioactivities of *Aristolochia* species such as insecticidal, anti-bacterial, anti-nociceptive, and anti-inflammatory effects (Kuo et al., 2011; Salome et al., 2020). Therefore, the strict supervision and accurate utilization of the *Aristolochia* species are important to implement the medicinal value.

Elucidating the relationships between species of genus *Aristolochia* is crucial for understanding and harnessing the medicinal properties of the different species. However, as a diverse genus with a large number of species distributed widely in geography, the circumscription and infrageneric classification of genus *Aristolochia* have been complicated and ambiguous. In the cladistic analysis based on morphological characters, many infrageneric taxa have been recognized by different authors (Ohi-Toma and Murata, 2016). For example, González et al. proposed that genus *Aristolochia* should be divided into three subgenera (*Aristolochia*, *Pararistolochia* and *Siphisia*), while Stevenson et al. indicated that the genus consisted of four genera in two subtribes Aristolochiinae and Isotrematinae (González and Rudall, 2003; Buchwalder et al., 2014; Ohi-Toma and Murata, 2016). Besides, in the *Flora of China*, it is also stated a controversy that some species of *Aristolochia* should be transferred to the genus *Isotrema* (Huang et al., 2003). Molecular markers are a reliable alternative that is independent of morphological feathers, enabling them to address the taxonomic challenges arising from the blurring morphological characters (Wu et al., 2020). Numerous molecular methods have been applied to *Aristolochia* and have advanced the understanding of the relationships of the genus *Aristolochia* (Wanke et al., 2006; González et al., 2014; Zhao et al., 2021). The phylogenetic trees produced with three gene sequences *rbcL*, *phyA* and *matk* of *Aristolochia* supported that *Aristolochia* was composed of two lineages corresponding to Aristolochiinae and Isotrematinae, respectively (Ohi-Toma et al., 2006). Based on the combined analysis using two plastid genic spacers (*rps16-trnK* and *petB-petD*) and two nuclear genes (*phyA* and ITS2), the phylogeny construction results confirmed that genus *Aristolochia* was divided into two well-supported clades representing subtribe Aristolochiinae and Isotrematinae, and Zhu et al. suggested *Aristolochia* subgenus *Siphisia* should be treated as an independent genus *Isotrema* (Zhu et al., 2019a). However, the results of different studies are not completely consistent, and the taxonomic systems of *Aristolochia* are still controversial (Wanke et al., 2006; Buchwalder et al., 2014; Ohi-Toma and Murata, 2016). With the new values and species of *Aristolochia* gradually published, effective methods to resolve the phylogenetic relationships and assess the previous classification of *Aristolochia* species are urgently needed (Yang et al., 2018; Luo et al., 2020).

With the rapid development of next-generation DNA sequencing (NGS) technologies, obtaining a complete plastome sequence has become a laboratory routine (Shi et al., 2019). The complete chloroplast (cp) genomes, as the important organelle DNA in plants, are characterized by a large size, containing richer variant site information to be an attractive tool for phylogenetic studies of plants (Niu et al., 2018). Compared with the phylogenetic analysis based on the limited phylogenetic information provided by short fragments of nuclear and cp DNA, the cp genome has significant advantages in phylogenetic resolution, particularly at low taxonomic levels (Parks et al., 2009; Wilkinson et al., 2017; Wu et al., 2021). For example, plastid genome data provided strong support for the sister relationship of sect. *Macroceras* and sect. *Diphyllon* of the genus *Epimedium* (Guo et al., 2022). In recent years, several cp genomes of *Aristolochia* have been reported (Kim and Lim, 2019; Zhao et al., 2021). The molecular structure and phylogenetic analyses of cp genomes of *Aristolochia debilis* and *Aristolochia contorta* revealed a close phylogenetic relationship with Piperaceae, Laurales, and Magnoliales (Zhou et al., 2017). Nevertheless, the compared analysis of multiple *Aristolochia* chloroplasts is still deficient, which is unable to comprehensively illustrate the intricate phylogenetic relationships and systematic evolution of *Aristolochia*.

In this study, we reported eleven complete *Aristolochia* cp genomes including five of subgenera *Siphisia* (*A. fulvicoma, A. hainanensis, A. griffithii, A. kwangsiensis* and *A. dabieshanensis*), three in subgenera *Aristolochia* (*A. tagala*, *A. debilis, A. tubiflora*) and another three species (*A. gigantea*, *A. littoralis, A. neolongifolia*) with unclear subgenera information. The comparative genomic analyses were conducted to explore the features and structural differentiation of the sequences. Analysis of simple sequence repeats (SSRs) could screen out potential molecular polymorphic markers for analyzing the genetic diversity and structure of *Aristolochia* populations in the future. Highly variable regions would provide candidate DNA barcodes for further studying *Aristolochia* species identification. Phylogenetic analysis performed by constructing phylogenetic trees enabled to reveal the interspecific relationship of *Aristolochia* species. This study enriched the valuable complete cp genome resources of *Aristolochia* and will contribute to further research on the identification and phylogenetic relationships within the species of the genus *Aristolochia*.

2 Materials and methods

2.1 Taxon sampling and DNA extraction

Eleven species of *Aristolochia* were newly collected from the Hainan, Yunnan, Xizang, Guizhou, and Hubei Provinces of China (Supplementary Table 1). Thereinto, referring to the *Flora of China* (http://www.iplant.cn/), five species (*A. fulvicoma, A. hainanensis, A. griffithii, A. kwangsiensis* and *A. dabieshanensis*) were divided into subgenus *Siphisia*, and other three species of *A. tagala*, *A. debilis* and *A. tubiflora* were recorded in subgenus *Aristolochia*. Besides, three species without taxonomic information on subgenus, *A. gigantea*, *A. littoralis*, and *A. neolongifolia*, were collected to explore the phylogeny. The 11 individuals were frozen at -80°C and the total genomic DNA was isolated from fresh leaves using the Plant Genomic DNA Kit (TIANGEN, Beijing, China) by the manufacturer's instructions. DNA integrity was examined by electrophoresis in 1% (w/v) agarose gel and their concentration was measured using a NanoDrop 2000C spectrophotometer (Thermo Scientific; Waltham, MA, USA).

2.2 DNA sequencing, assembly and annotation

The quantified DNA was used to construct shotgun libraries with insert sizes of 300~500bp and a paired-end library was constructed by TruSeq™ Nano DNA Sample Prep Kit (Illumina, San Diego, CA, USA). Then paired-end sequencing was performed to obtain 150 bp sequences at both ends of each read according to the manufacturer's manual for the Illumina NovaSeq platform (Illumina, San Diego, CA, USA). Low-quality regions in the original data were trimmed using the software Trimmomatic (Bolger et al., 2014). Then the clean cp reads were screened and compared with the *Aristolochia* sequences published at the National Centre for Biological Information. SOAPdenovo 2 was used to splice the extracted reads into several contigs (Luo et al., 2012). The assembled contigs were connected to cp genome sequences by using the NOVOPlasty (Dierckxsens et al., 2017), and gaps were filled by the GapCloser module in SOAP package. Lastly, the genes, introns and boundaries of coding regions were compared with reference sequences, *A. debilis* (NC036153), and assembled into complete cp genomes. Genome annotation was performed referring to the complete cp genomes of *Aristolochia* and corrected manually. All of the annotated genomes were deposited in GenBank with the accession numbers listed in Supplementary Table 1.

2.3 Genome structure analyses

Chloroplast circular maps were drawn in OGDRAW v1.3.1 (http://ogdraw.mpimp-golm.mpg.de/) according to the adjusted genome annotation. The GC content was analyzed using MEGA (Tamura et al., 2013). The SSRs were identified by MISA software (Beier et al., 2017) with the thresholds of 10, 5, 4, 3, 3, and 3 repeat units for mono-, di-, tri-, tetra-, penta-, and hexanucleotides, respectively. To identify the long repeat motifs, REPuter (Kurtz et al., 2001) was used to locate direct, reverse, complementary and palindromic sequences, with a minimum repeat size of 30bp and Hamming distance of 3. Statistical analysis was accomplished by GraphPad Prism (GraphPad Sofware, La Jolla, CA, USA).

2.4 Comparative and phylogenetic analyses

The whole-genome alignment for the 11 *Aristolochia* cp genomes was performed and plotted using mVISTA software (Dubchak and Ryaboy, 2006). Comparison of boundaries of the large single-copy (LSC), small single-copy (SSC) and two inverted repeats (IR) regions was analyzed using IRscope (Amiryousefi et al., 2018). The nucleotide diversity (Pi) of shared genes and intergenic spacers was calculated using DnaSP (Librado and Rozas, 2009). The cp genomes of the 11 *Aristolochia* species together with those *Aristolochia* species available in NCBI, which were *A. bracteolata* (MT130705), *A. tagala* (NC041455), *A. debilis* (NC036153), *A. delavayi* (MW413320), *A. kaempferi* (NC041452), *A. mollissima* (NC041457), *A. kunmingensis* (NC041451), *A. moupinensis* (NC041454), *A. kwangsiensis* (NC052833) and *A. macrophylla* (NC041453), were used for phylogenetic analyses. The cp genomes of *Asarum pulchellum* (MZ440306) and *Piper kadsura* (NC027941) were included as the outgroup to root the tree. Considering the better-supported trees yielded by protein-coding data sets, a total of 72 protein-coding genes which were shared by these species were extracted to perform ML analysis using PhyloSuite software (Zhang et al., 2020; Guo et al., 2022). The maximum-likelihood (ML) analysis was performed based on the generated data using IQ-TREE with 1000 bootstrap replicates (Nguyen et al., 2015).

3 Results

3.1 Structure features of *Aristolochia* plastomes

The complete cp genomes of 11 *Aristolochia* species were all typical quadripartite structures with the total length from 159,375 bp (*A. tagala*) to 160,626 bp (*A. tubiflora*) (Figure 1; Table 1). The consisted LSC region (88914-90251 bp) and SSC region (19311-19917 bp) were separated by two inverted repeat (IR) regions (50350-51396 bp) (Table 1). The total number of unique genes annotated is from 130 to 131, comprising 85 protein-coding genes (CDS), 37-38 tRNA and 8 rRNA genes (Table 1). GC contents of the plastomes of 11 *Aristolochia* species ranged slightly from 38.3% to 38.8%, and the GC contents of the four regions were not balanced. The IR regions had the highest GC content (43.4-43.6%), followed by the LSC regions (36.6-37.2%) and the SSC regions (32.8-33.8%) (Supplementary Table 2). The cumulative length of CDS ranged from 77,466 (*A. littoralis*) to 79,074 bp (*A. gigantea*) and the GC contents were 38.9% to 39.2% (Table 1; Supplementary Table 2). Moreover, the GC% content of the first position was higher compared to those of the second and third positions (Supplementary Table 2).

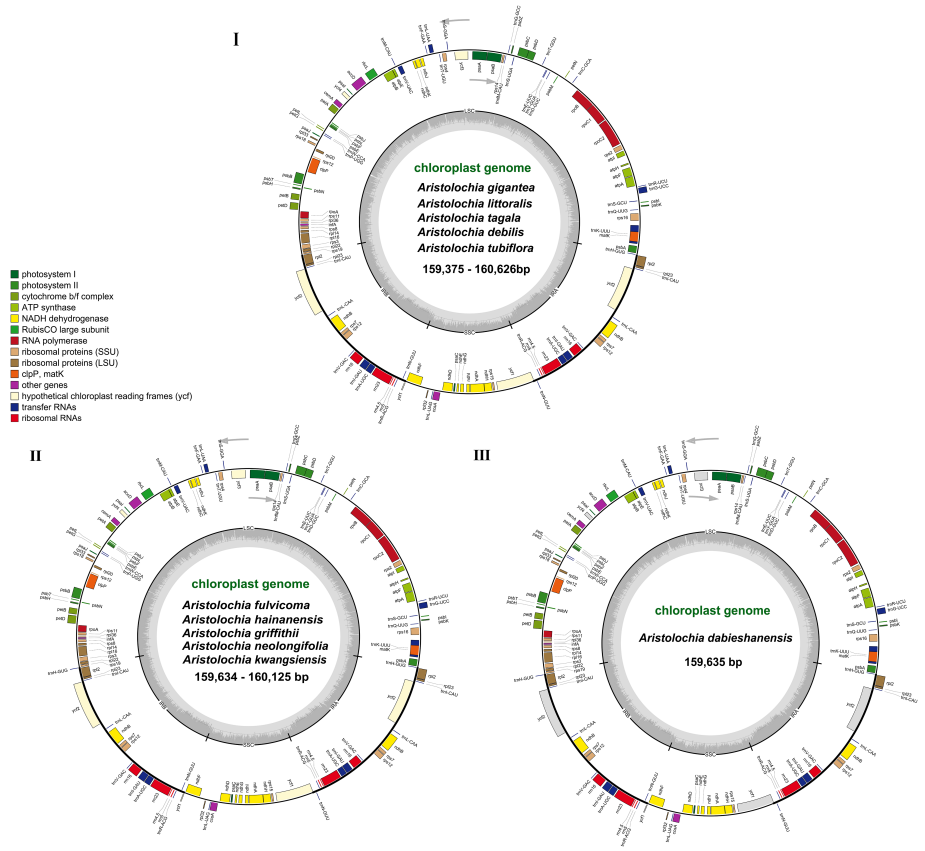


FIGURE 1
Gene maps of the complete cp genome of 11 species of Aristolochia. Three types of cp genome of **(I)** *A. gigantea, A. littoralis, A. tagala, A. debilis* and *A. tubiflora*. **(II)** *A. fulvicoma, A. hainanensis, A. griffithii, A. neolongifolia* and *A. kwangsiensis*. **(III)** *A. dabieshanensis*. Genes on the inside of the circle are transcribed clockwise, while those outside are transcribed counter clockwise. The darker gray in the inner circle corresponds to the GC content, whereas the lighter gray corresponds to AT content.

3.2 Repeat structure and simple sequence repeats analyses

A total of 817 repeats were identified in 11 *Aristolochia* species, including 288 reverse repeats, 260 palindromic repeats, 175 complement repeats, and 94 direct repeats (Supplementary Table 3). For each *Aristolochia* species, the number of repeat sequences varied greatly. *A. littoralis* had the largest number of repeats (168), while *A. tagala* had the smallest number of repeats (42). Four types of repeating motifs were detected in all 11 species (Figure 2A; Supplementary Table 3). The length of these repeats was mainly concentrated in 30-49 bp. Repeats with a length of ≥50bp only existed in *A. gigantea* and *A. littoralis* (Figure 2B; Supplementary Table 4).

Six kinds of SSRs were screened in the cp genomes of 11 *Aristolochia* species. The number of SSRs identified in 11 *Aristolochia* plastomes ranged from 99 in *A. kwangsiensis* to 161 in *A. gigantea* (Supplementary Table 5). In these SSRs, mono-nucleotide repeats were the largest in number, which accounted for the percent of 59.57%-72.61% in all types of SSRs (Figure 3A; Supplementary Table 5). The base composition of the repeating motifs had a certain base preference, mainly the repeating motifs rich in A-T

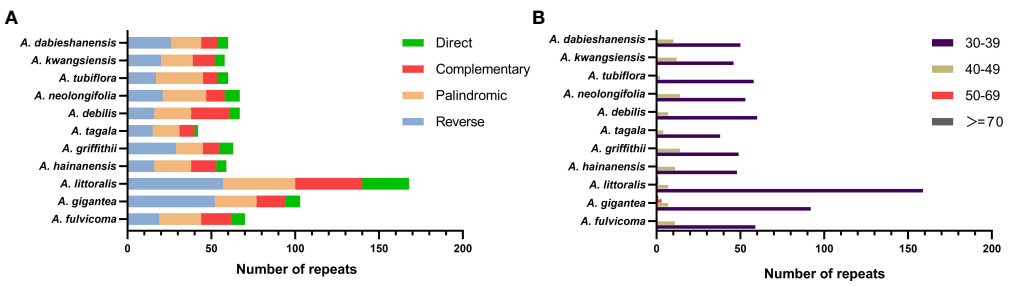


FIGURE 2
Long repeat sequences in the chloroplast genome of 11 species of *Aristolochia*. **(A)** Frequency of four types of repeats; **(B)** Length of repeat sequences.

(Supplementary Table 5). Eleven species all contained six kinds of repeat except for *A. kwangsiensis* and *A. dabieshanensis* which were without Hexa (Figure 2A; Supplementary Table 5). Regarding the SSRs distribution, these SSRs were mainly found in the LSC regions (Figure 3B; Supplementary Table 6).

3.3 Comparative genomic divergence and hotspots regions

Divergence hotspots are important for discovering DNA markers and barcodes in species identification (Kong et al., 2021). In this study, the cp genomes of 11 species of *Aristolochia* were compared using mVISTA with the *A. debilis* genome as the reference genome. Overall, the comparative genomic analysis revealed that the 11 *Aristolochia* cp genomes were relatively conserved. Most variations are discovered in the conserved noncoding sequences, and only a few in coding genes, such as *accD*, *ndhF* and *ycf1* (Figure 4). The results indicated that the coding-gene sequences were more conserved than the noncoding sequences. Moreover, the nucleic acid variation analyses showed the intergenic spacers had more polymorphisms (average Pi=0. 04049) than the gene regions (average Pi=0.01546) (Figure 5). The highly variable regions comprised the genes regions: *clpP*, *matK*, *ndhF*, *psbT*, *rps16*, *trnK*-UUU (Pi>0.035). Among the six highly variable regions, five regions *clpP*, *matK*, *psbT*, *rps16*, and *trnK*-UUU were located in the LSC, and *ndhF* was located in the SSC. The intergenic spacer regions with high variations were screened as follows: *ccsA-ndhD*, *psbZ-trnG*-GCC, *rpl33-rps18*, *rps16-trnQ*-UUG, *trnS*-GCU-*trnG*-UCC (Pi>0.060). Among the five highly variable regions, four regions, *rpl33-rps18*, *rps16-trnQ*-UUG, *psbZ-trnG*-GCC, and t*rnS*-GCU-*trnG*-UCC were located in the LSC, and *ccsA-ndhD* were located in the SSC. It was confirmed that the variations in the LSC and SSC regions were remarkably higher than those in the IR regions of cp.

3.4 Phylogenetic analyses

Chloroplast genomes play an important role in phylogenetic studies, and it is necessary to solve complex evolutionary relationships (Zhang et al., 2011). In our study, to obtain a more accurate analysis of the *Aristolochia* phylogeny, available *Aristolochia* genomes downloaded from NCBI were also included to construct the phylogenetic tree. A total of eighteen *Aristolochia* species were contained, and *Asarum pulchellum* and *Piper kadsura* served as the outgroup (Figure 6). Phylogenetic analyses using the ML method and sequences of 72 CDS strongly supported the identification of two clades among *Aristolochia* species, and they corresponded to subgenus *Aristolochia* (Clade A) and subgenus *Siphisia* (Clade B), as classified in *Flora of China* (Huang et al., 2003). Within the subgenus *Aristolochia*, *A. gigantea* and *A. littoralis* formed a monophyletic cluster, which was a sister to the other five *Aristolochia* species (*A. bracteolate*, *A. tagala*, *A. delavayi*, *A. tubiflora* and *A. debilis*). In subgenus *Siphisia*, *A. macrophylla* diverged first. Then *A. griffithii* showed a sister relationship with remaining *Siphisia* species. The monophyletic cluster comprising *A. fulvicoma*, *A. kwangsiensis* and *A. hainanensis* was a sister to the cluster composed of *A. kunmingensis*, *A. neolongifolia* and *A.moupinensis*, and both were sister to the other three species (*A. kaempferi*, *A. mollissima* and *A. dabieshanensis*).

3.5 IR expansion and contraction investigation

The boundaries of IR region are hot spots for gene duplications or deletions (Yue et al., 2008). In this study, the expansion and contraction of the IR region in 11 *Aristolochia* cp genomes were analyzed. Results showed that all *Aristolochia* plastomes have the SSC/IRb boundary within the pseudogene (ψ) *ycf1* gene and the SSC/IRa border within the *ycf1* gene except the *A. tagala* which between the *ycf1* and gene *trnN* (Figure 7). However, there were some differences in the IR/LSC border area, and three types of plastomes were characterized by IR/LSC boundary variation (Figures 1, 7). *A. debilis*, *A. tubiflora*, *A. tagala*, *A. gigantea* and *A. littorali* were grouped together in Clade A and classified as Type I, because the LSC-IRb border of cp genomes was located within the genic spacer of *rps19-rpl2* as well as the LSC-IRa border was located within the gene *trnH*. Type II and III corresponded to Clade B which contained one more repeat of *trnH*-GUG in the IRb region and the LSC-IRb border was located within the gene *rps19* (*A. fulvicoma, A. hainanensis, A. griffithii, A. neolongifolia* and *A. kwangsiensis*) or the genic spacer of *rps19-trnH* (*A. dabieshanensis*). Besides, the IRa regions of Type II and III had slightly expanded, resulting in *trnH* being located in the

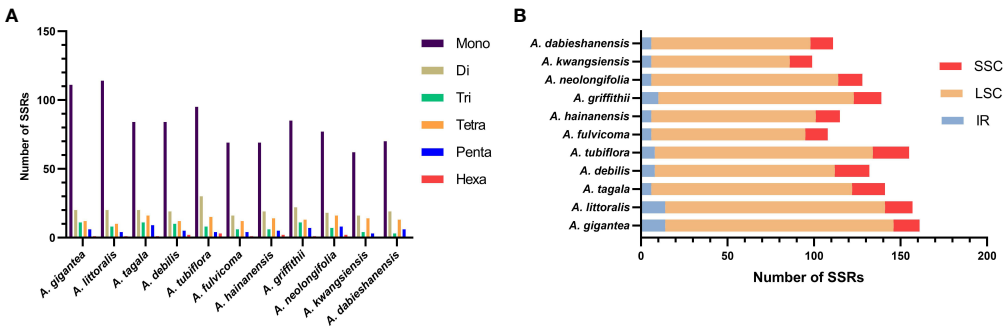


FIGURE 3
Amounts and distribution of SSRs in the chloroplast genome of 11 species of *Aristolochia*. **(A)** Amounts of the SSRs; **(B)** Distribution of SSRs.

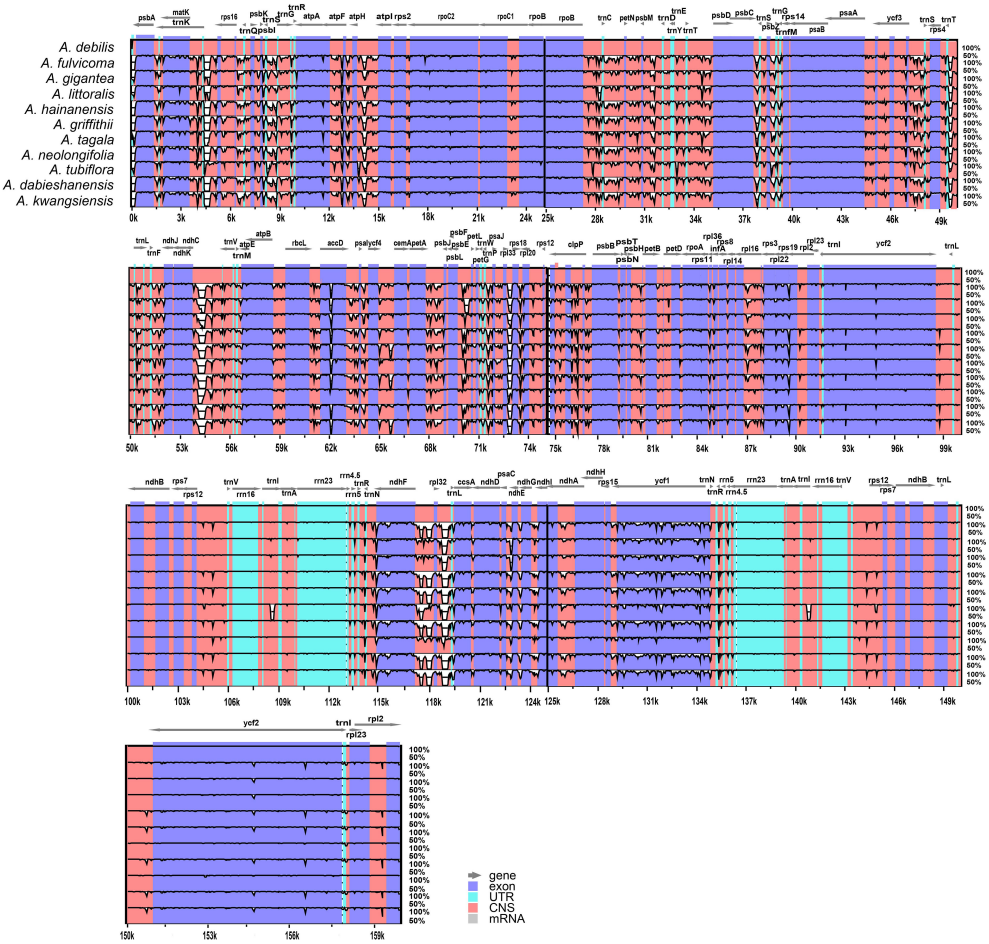


FIGURE 4
Sequence identity plot comparing the four chloroplast genomes of species of *Aristolochia* with *Aristolochia debilis* as a reference using mVISTA. Gray arrows and thick black lines above the alignment indicate genes with their orientation. Purple bars represent exons, blue bars represent UTRs, and red bars represent noncoding sequences (CNS). Y-scale represents the percent identity ranging from 50% to 100%.

IR region, and the IRa/LSC border was located in the genic spacer of *trnH-psbA*. Compared with the type I, the IR region of the remaining two types of plastomes expanded approximately 0.4-1.0 k (Table 1; Figure 7).

4 Discussion

In this study, we reported eleven sequenced complete cp genomes of *Aristolochia* (Figure 1; Table 1). Our comparative analyses indicated that the overall eleven cp genomes showed a highly conserved feature in terms of structures. The GC content of eleven *Aristolochia* plants ranged from 38.3% to 38.8%, which was the same as that previously reported *Aristolochia* plastome (Zhou et al., 2017; Li et al., 2019). Besides, the IR regions had the highest GC content among the four regions of these *Aristolochia* species, which was consistent with most other angiosperms (Wu et al., 2020). SSRs, also known as cp microsatellites, were short tandem repeat sequences consisting of 1-6 bp nucleotides as repeating units. In all types of SSRs in this study, A or T repeats accounted for the majority, and mono-nucleotides were the predominant type. The richness of A/T in cp genomes can be explained by the easier strand separation for increasing the slipped-strand mispairing as compared to GC/CG and other tracts (George et al., 2015). Widely distributed SSRs in cp genomes provide the available molecular markers for the species of interest or closely related species (Varshney et al., 2005; Vu et al., 2020). In orchids, SSR markers were developed for recognizing valuable plants, investigating intraspecific genetic variation and reconstructing phylogeographic patterns (Tsa et al., 2014). The SSRs detected in the *Aristolochia* species were of great significance for the phylogenetic research and classification of *Aristolochia* plants. Additionally, four types of long repeat sequences were all identified in 11 *Aristolochia* species including direct, reverse, complement and palindromic. Most of the repeats were reversed and palindromic. These long repeat sequences were not only abundant in mutations but also very important in phylogenetic analyses (Wu et al., 2021). All the identified repeats in this study may be useful for the population genetics studies of these 11 species in the future.

Although the cp genomes among 11 *Aristolochia* species have a highly conserved feature, there were some small changes presented

10.3389/fpls.2023.1119041

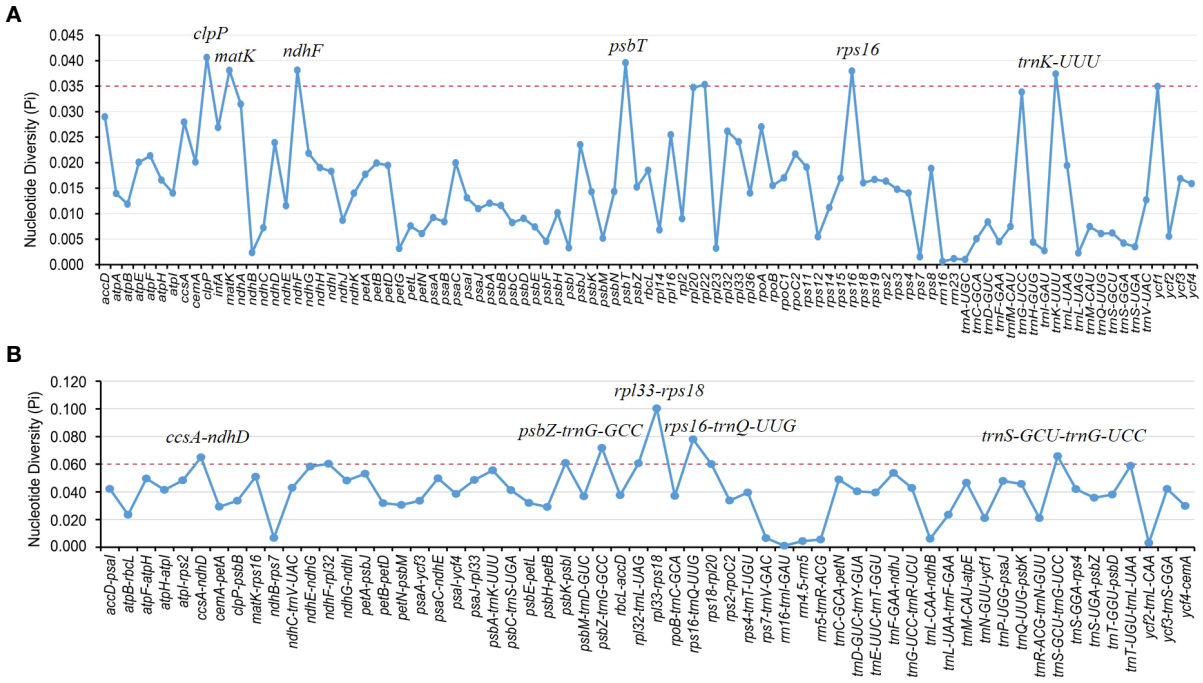


FIGURE 5
Nucleotide diversity (Pi) of shared various regions in 11 *Aristolochia* species chloroplast genomes. **(A)** Pi values in the genes regions. **(B)** Pi values in the intergenic spacers regions.

among these species on the boundary between the IR and LSC regions. Plastid genomes have been divided into three types according to the boundary of LSC/IR, which has a certain relationship corresponding to the clades of two subgenera *Aristolochia* and *Siphisia*. By comparing the length variation of IR, LSC and SSC regions among these cp genomes, we also found that the IR region of the plastomes of subgenus *Siphisia* expanded approximately 0.4-1.0 k to the LCS region compared to the subgenus *Aristolochia* (Figure 7). The expansions and contractions of the boundaries of the IR regions are considered to be the main reason for the size change of cp (Zhang et al., 2016). Besides, the deletion of one copy of *trnH*-GUG gene was observed in subgenus *Aristolochia* species, which resulted in the total of 37 tRNA genes in the species of subgenus *Aristolochia* and 38 in *Siphisia* (Table 1). A previous study also reported the loss of the *trnH*-GUG genes was one of the major differences between the plastomes of the two subgenera

TABLE 1 Genome structure of complete chloroplast (cp) genomes of *Aristolochia* species.

Species	Genome Size (bp)	LSC (bp)	IR (bp)	SSC (bp)	CDS (bp)	Number of genes	rRNA	tRNA	Protein-coding genes	GC %	Genome type
A. gigantea	160594	90230	50854	19510	79074	130	8	37	85	38.4	I
A. littoralis	160610	90251	50886	19473	77466	130	8	37	85	38.4	I
A. tagala	159375	89441	50522	19412	78603	130	8	37	85	38.5	I
A. debilis	159812	89627	50350	19835	78708	130	8	37	85	38.3	I
A. tubiflora	160626	89945	50764	19917	78801	130	8	37	85	38.4	I
A. fulvicoma	159806	89200	51282	19324	78966	131	8	38	85	38.8	II
A. hainanensis	159672	89011	51340	19321	78954	131	8	38	85	38.8	II
A. griffithii	160125	89404	51386	19335	78921	131	8	38	85	38.7	II
A. neolongifolia	159634	88914	51396	19324	78948	131	8	38	85	38.8	II
A. kwangsiensis	159734	89069	51354	19311	78951	131	8	38	85	38.8	II
A. dabieshanensis	159635	88933	51362	19340	78942	131	8	38	85	38.8	III

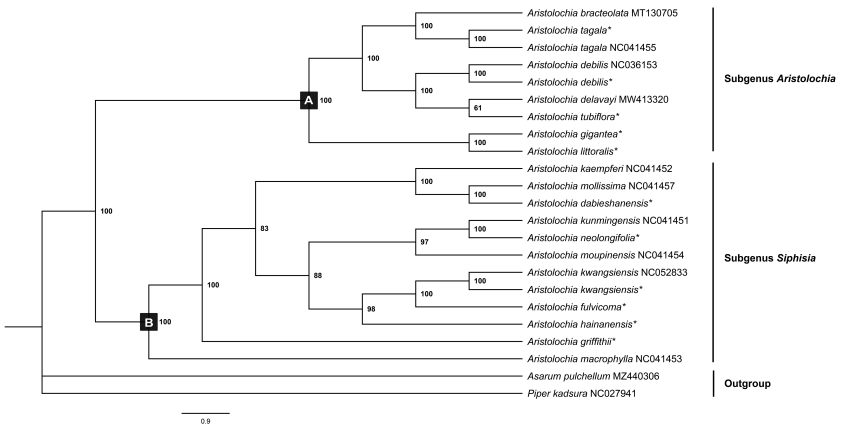


FIGURE 6
Phylogenetic tree inferred from the CDS of the 72 protein-coding genes of *Aristolochia* species using Maximum Likelihood (ML) method. The numbers near by nodes are values for bootstrap support. Species with newly sequenced chloroplast genomes are marked with asterisks.

Siphisia and *Aristolochia* (Li et al., 2019). These sequence variations might be the result of boundary contraction and expansion between the LSC/IR regions in plants (Wang et al., 2022). Plastid genomes have the characteristics of high conservation and slow evolutionary rate, thus the special characteristics presented in their structure are often phylogenetically informative (Pascual-Diaz et al., 2021). In general, broad sampling and more evidence from the genomes will be necessary for the further understanding of the interspecies relationships of *Aristolochia*.

Species of *Aristolochia* are controversially officinal and strictly forbidden in the present. The identification of *Aristolochia* species is important to supervise the abuse and protect customer safety. Morphological evidence is a conventional method for plant classification and identification. However, morphological traits are easily affected by the natural environment and artificial treatment, which hardly meet the requirements of detection in practical application. DNA studies can achieve the accurate authentication of similar species within a genus based on reliable molecular evidence. Numerous DNA regions, such as the nuclear genes ITS2, and cp genes *matK*, *rbcL*, *trnH-psbA* and *trnL-trnF*, have been applied to the identification of *Aristolochia* species (Li et al., 2014; Dechbumroong et al., 2018). However, multiple primers were required to achieve the authentication of different *Aristolochia* species, and the existence of long sequence deletions or poly-A/T sequences also resulted in the difficulty of sequencing analysis (Wu et al., 2015). In this study, the results of mVISTA analysis suggested that the hypervariable intergenic regions were mostly distributed in the non-coding regions, and rarely in coding genes. Moreover, our comparative results have shown that the used cp markers appeared to be relatively low in nucleotide diversity, which may be insufficient to

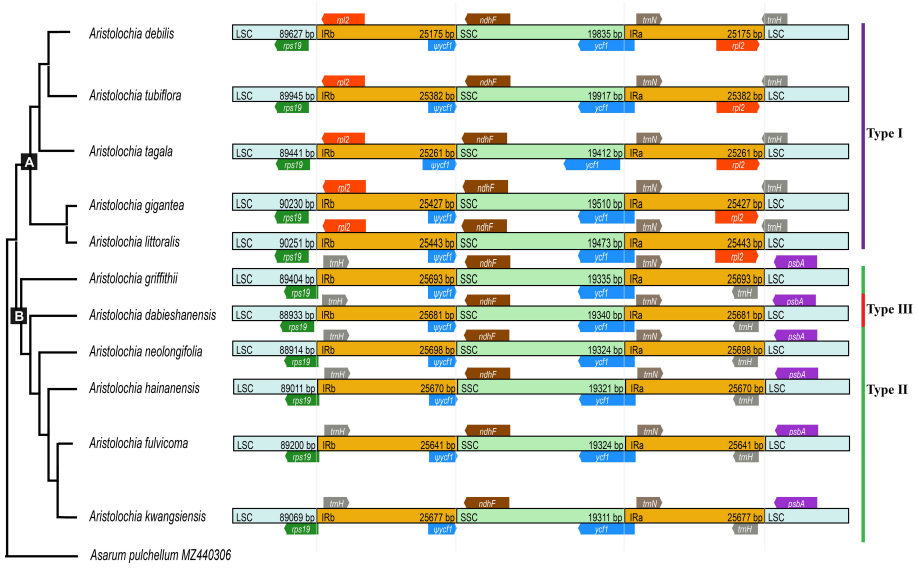


FIGURE 7
Comparison of boundaries of the LSC, SSC, and IR regions among 11 *Aristolochia* plastomes.

distinguish the species within genus *Aristolochia*. Thus, to achieve better species resolution, future molecular markers can focus on the more variable regions of the cp genomes, such as *clpP*, *psbT*, *rps16*, *ycf1* and *rpl33-rps18* (Chang et al., 2021).

With increasing taxon samples of *Aristolochia* species, our phylogenetic analyses of cp genome sequences have substantially improved the phylogenetic resolution and provided robust inference of the intraspecific relationships. In the current study, phylogenetic trees of the genus *Aristolochia* were constructed based on CDS sequences from a total of 18 *Aristolochia* species, including eleven species we sequenced and other seven downloaded from NCBI. Regarding the division of genus *Aristolochia*, our phylogenetic analyses have confirmed the division of two clades representing the species of subgenus *Aristolochia* and *Siphisia*, respectively. This cp phylogeny concurs well with previously published phylogenetic trees based on several nuclear/plastid regions (Zhu et al., 2019a). Compared with the phylogenetic results, it is further confirmed that the species clustered in subgenus *Siphisia* also could be corresponded with the *Isotrema* species, which is consistent with the classification based on the morphological characteristics, number of chromosomes and molecular data (Huang et al., 2003; Ohi-Toma et al., 2006; Zhu et al., 2019b). Our result provided stronger support that the subgenus *Siphisia* was clustered as an independent clade, and may contribute to the reinstatement of *Isotrema* as a new generic delimitation of *Aristolochia* subgenus *Siphisia*. In general, the phylogenetic tree conducted in this study demonstrated that the cp genomes can be used as essential evidence to resolve the intergeneric and interspecies relationships within genus *Aristolochia*.

5 Conclusion

In this study, the complete cp genomes of eleven species of genus *Aristolochia* were sequenced and compared. All of these cp genomes were obvious quadripartite structures and comparatively conserved on the length, GC content and gene content. The high variations were mostly found in LCS and SSC regions, and variable regions could serve as potential markers for species identification. Phylogenetic results indicated that the genus *Aristolochia* was composed of two main clades, corresponding to the division of subgenus *Siphisia* and subgenus *Aristolochia*. Moreover, combined with the analyses of IR/LSC boundaries, a whole duplication of *trnH*-GUG gene was observed in subgenus *Siphisia*, and it may be associated with the expansion of its IR region. In conclusion, this study provides an important foundation for species identification and valuable insight into the phylogenetic relationships of the *Aristolochia*.

Data availability statement

The data presented in the study are deposited in the NCBI repository (https://www.ncbi.nlm.nih.gov/), and the accession numbers were OP895634, OP925753, OP950686-OP950694.

Author contributions

JH, XB, GW, YR and YS conceived and designed the study. XB, GW collected and analyzed the data. XB, YR and YS wrote the manuscript. All authors have directly contributed to this manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study is supported by the National Key Research and Development Program of China (2019YFC1604701) and CAMS Innovation Fund for Medical Sciences, China (CIFMS, 2021-I2M-1-071).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1119041/full#supplementary-material

References

Amiryousefi, A., Hyvonen, J., and Poczai, P. (2018). IRscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics* 34 (17), 3030–3031. doi: 10.1093/bioinformatics/bty220

Beier, S., Thiel, T., Muench, T., Scholz, U., and Mascher, M. (2017). MISA-web: a web server for microsatellite prediction. *Bioinformatics* 33 (16), 2583–2585. doi: 10.1093/bioinformatics/btx198

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinf. (Oxford England)* 30 (15), 2114–2120. doi: 10.1093/bioinformatics/btu170

Buchwalder, K., Samain, M.-S., Sankowsky, G., Neinhuis, C., and Wanke, S. (2014). Nomenclatural updates of aristolochia subgenus pararistolochia (Aristolochiaceae). *Aust. Systematic Bot.* 27 (1), 48–55. doi: 10.1071/sb13042

Chang, H., Zhang, L., Xie, H., Liu, J., Xi, Z., and Xu, X. (2021). The conservation of chloroplast genome structure and improved resolution of infrafamilial relationships of crassulaceae. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.631884

Dechbumroong, P., Aumnouypol, S., Denduangboripant, J., and Sukrong, S. (2018). DNA Barcoding of aristolochia plants and development of species-specific multiplex PCR to aid HPTLC in ascertainment of aristolochia herbal materials. *PLoS One* 13 (8), e0202625. doi: 10.1371/journal.pone.0202625

Dierckxsens, N., Mardulyn, P., and Smits, G. (2017). NOVOPlasty: *de novo* assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* 45 (4), e18. doi: 10.1093/nar/gkw955

Dubchak, I., and Ryaboy, D. V. (2006). VISTA family of computational tools for comparative analysis of DNA sequences and whole genomes. *Methods Mol. Biol. (Clifton N.J.)* 338, 69–89.

George, B., Bhatt, B. S., Awasthi, M., George, B., and Singh, A. K. (2015). Comparative analysis of microsatellites in chloroplast genomes of lower and higher plants. *Curr. Genet.* 61 (4), 665–677. doi: 10.1007/s00294-015-0495-9

González, F., and Rudall, P. J. (2003). Structure and development of the ovule and seed in aristolochiaceae, with particular reference to saruma. *Plant Systematics Evol.* 241 (3-4), 223–244. doi: 10.1007/s00606-003-0050-x

González, F., Wagner, S. T., Salomo, K., Symmank, L., Samain, M.-S., Isnard, S., et al. (2014). Present trans-pacific disjunct distribution of aristolochia subgenus isotrema (Aristolochiaceae) was shaped by dispersal, vicariance and extinction. *J. Biogeography* 41 (2), 380–391. doi: 10.1111/jbi.12198

Guo, M., Pang, X., Xu, Y., Jiang, W., Liao, B., Yu, J., et al. (2022). Plastid genome data provide new insights into the phylogeny and evolution of the genus epimedium. *J. Adv. Res.* 36, 175–185. doi: 10.1016/j.jare.2021.06.020

Huang, S., Kelly, L., and Gilbert, M. (2003). Aristolochiaceae. *Flora China* 5, 246–269.

Jelaković, B., Dika, Ž., Arlt, V. M., Stiborova, M., Pavlović, N. M., Nikolić, J., et al. (2019). Balkan Endemic nephropathy and the causative role of aristolochic acid. *Semin. Nephrol.* 39 (3), 284–296. doi: 10.1016/j.semnephrol.2019.02.007

Ji, H., Hu, J., Zhang, G., Song, J., Zhou, X., and Guo, D. (2021). Aristolochic acid nephropathy: A scientometric analysis of literature published from 1971 to 2019. *Med. (Baltimore)* 100 (27), e26510. doi: 10.1097/MD.0000000000026510

Kim, K., and Lim, C. E. (2019). The complete chloroplast genome sequence of aristolochia manshuriensis kom. (Aristolochiaceae). *Mitochondrial DNA B Resour* 4 (2), 3515–3516. doi: 10.1080/23802359.2019.1675484

Kong, B. L.-H., Park, H.-S., Lau, T.-W. D., Lin, Z., Yang, T.-J., and Shaw, P.-C. (2021). Comparative analysis and phylogenetic investigation of Hong Kong ilex chloroplast genomes. *Sci. Rep.* 11 (1), 5153. doi: 10.1038/s41598-021-84705-9

Kuo, P.-C., Li, Y.-C., and Wu, T.-S. (2011). Chemical constituents and pharmacology of the aristolochia species. *J. Traditional Complementary Med.* 2 (4), 249–266.

Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* 29 (22), 4633–4642. doi: 10.1093/nar/29.22.4633

Li, M., Au, K.-Y., Lam, H., Cheng, L., But, P. P.-H., and Shaw, P.-C. (2014). Molecular identification and cytotoxicity study of herbal medicinal materials that are confused by aristolochia herbs. *Food Chem.* 147, 332–339. doi: 10.1016/j.foodchem.2013.09.146

Li, X., Zuo, Y., Zhu, X., Liao, S., and Ma, J. (2019). Complete chloroplast genomes and comparative analysis of sequences evolution among seven aristolochia (Aristolochiaceae) medicinal species. *Int. J. Mol. Sci.* 20 (5), 1–27. doi: 10.3390/ijms20051045

Librado, P., and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25 (11), 1451–1452. doi: 10.1093/bioinformatics/btp187

Luo, R. B., Liu, B. H., Xie, Y. L., Li, Z. Y., Huang, W. H., Yuan, J. Y., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* 1, 1–6. doi: 10.1186/2047-217x-1-18

Luo, Y. J., Ni, S. D., Jiang, Q., Huang, B. G., Liu, Y., and Huang, Y. S. (2020). Aristolochia yachangensis, a new species of aristolochiaceae from limestone areas in guangxi, China. *PhytoKeys* 153, 49–61. doi: 10.3897/phytokeys.153.52796

Maggini, V., Menniti-Ippolito, F., and Firenzuoli, F. (2018). Aristolochia, a nephrotoxic herb, still surfs on the web, 15 years later. *Intern. Emerg. Med.* 13 (5), 811–813. doi: 10.1007/s11739-018-1813-2

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32 (1), 268–274. doi: 10.1093/molbev/msu300

Niu, Y. T., Jabbour, F., Barrett, R. L., Ye, J. F., Zhang, Z. Z., Lu, K. Q., et al. (2018). Combining complete chloroplast genome sequences with target loci data and morphology to resolve species limits in triplostegia (Caprifoliaceae). *Mol. Phylogenet Evol.* 129, 15–26. doi: 10.1016/j.ympev.2018.07.013

Ohi-Toma, T., and Murata, J. (2016). Nomenclature of isotrema, siphisia, and endodeca, and their related infrageneric taxa of aristolochia (Aristolochiaceae). *Taxon* 65 (1), 152–157. doi: 10.12705/651.11

Ohi-Toma, T., Sugawara, T., Murata, H., Wanke, S., Neinhuis, C., and Murata, J. (2006). Molecular phylogeny of aristolochia sensu lato (Aristolochiaceae) based on sequences of rbcL, matK,and phyA genes, with special reference to differentiation of chromosome numbers. *Systematic Bot.* 31 ((3), 481–492.

Parks, M., Cronn, R., and Liston, A. (2009). Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biol.* 7, 84. doi: 10.1186/1741-7007-7-84

Pascual-Diaz, J. P., Garcia, S., and Vitales, D. (2021). Plastome diversity and phylogenomic relationships in asteraceae. *Plants (Basel)* 10 (12), 1–16. doi: 10.3390/plants10122699

Salome, D. D. C., Cordeiro, N. M., Valerio, T. S., Santos, D. A., Alves, P. B., Alviano, C. S., et al. (2020). Aristolochia trilobata: Identification of the anti-inflammatory and antinociceptive effects. *Biomedicines* 8 (5), 1–21. doi: 10.3390/biomedicines8050111

Shi, L., Chen, H., Jiang, M., Wang, L., Wu, X., Huang, L., et al. (2019). CPGAVAS2, an integrated plastome sequence annotator and analyzer. *Nucleic Acids Res.* 47 (W1), W65–W73. doi: 10.1093/nar/gkz345

Stefanovic, V., Toncheva, D., Atanasova, S., and Polenakovic, M. (2006). Etiology of Balkan endemic nephropathy and associated urothelial cancer. *Am. J. Nephrol.* 26 (1), 1–11. doi: 10.1159/000090705

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30 (12), 2725–2729. doi: 10.1093/molbev/mst197

Tomlinson, T., Fernandes, A., and XXXP.Grollman, P. (2020). Aristolochia herbs and iatrogenic disease: The case of portland's powders. *Yale J. OF Biol. AND Med.* 93, 355–363.

Tsa, C.-C., Wu, P.-Y., Kuo, C.-C., Huang, M.-C., Yu, S.-K., Hsu, T.-W., et al. (2014). Analysis of microsatellites in the vulnerable orchid gastrodia flavilabella: the development of microsatellite markers, and cross-species amplification in gastrodia. *Botanical Stud.* 55, 72.

Vanherweghem, J., Depierreux, M., Tieiemans, C., Abramowicz, D., Dratwa, M., Jadoul, J., et al. (1993). Rapidly progressive interstitial renal fibrosis in young women: association with slimming regimen including Chinese herbs. *Lancet* 341 (8842), 387–391.

Varshney, R. K., Graner, A., and Sorrells, M. E. (2005). Genic microsatellite markers in plants: features and applications. *Trends Biotechnol.* 23 (1), 48–55. doi: 10.1016/j.tibtech.2004.11.005

Vu, H. T., Tran, N., Nguyen, T. D., Vu, Q. L., Bui, M. H., Le, M. T., et al. (2020). Complete chloroplast genome of paphiopedilum delenatii and phylogenetic relationships among orchidaceae. *Plants (Basel)* 9 (1), 1–27. doi: 10.3390/plants9010061

Wang, Y., Wen, F., Hong, X., Li, Z., Mi, Y., and Zhao, B. (2022). Comparative chloroplast genome analyses of paraboea (Gesneriaceae): Insights into adaptive evolution and phylogenetic analysis. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.1019831

Wanke, S., Gonza´lez, F., and Neinhuis, C. (2006). Systematics of pipevines: combining morphological and fast-evolving molecular characters to investigate the relationships within subfamliy aristolochioideae (Aristolochiaceae). *Int. J. Plant Sci.* 167 (6), 1215–1227.

Wilkinson, M. J., Szabo, C., Ford, C. S., Yarom, Y., Croxford, A. E., Camp, A., et al. (2017). Replacing Sanger with next generation sequencing to improve coverage and quality of reference DNA barcodes for plants. *Sci. Rep.* 7, 46040. doi: 10.1038/srep46040

Wu, L., Nie, L., Wang, Q., Xu, Z., Wang, Y., He, C., et al. (2021). Comparative and phylogenetic analyses of the chloroplast genomes of species of paeoniaceae. *Sci. Rep.* 11 (1), 1–10, 14643. doi: 10.1038/s41598-021-94137-0

Wu, L., Nie, L., Xu, Z., Li, P., Wang, Y., He, C., et al. (2020). Comparative and phylogenetic analysis of the complete chloroplast genomes of three paeonia section moutan species (Paeoniaceae). *Front. Genet.* 11. doi: 10.3389/fgene.2020.00980

Wu, L., Sun, W., Wang, B., Zhao, H., Li, Y., Cai, S., et al. (2015). An integrated system for identifying the hidden assassins in traditional medicines containing aristolochic acids. *Sci. Rep.* 5 (1), 1–10. doi: 10.1038/srep11318

Yang, B., Ding, H. B., Zhou, S. S., Zhu, X., Li, R., Maw, M. B., et al. (2018). Aristolochia sinoburmanica (Aristolochiaceae), a new species from north Myanmar. *PhytoKeys* 94), 13–22. doi: 10.3897/phytokeys.94.21557

Yue, F., Cui, L., dePamphilis, C. W., Moret, B. M., and Tang, J. (2008). Gene rearrangement analysis and ancestral order inference from chloroplast genomes with inverted repeat. *BMC Genomics* 9 Suppl 1, S25. doi: 10.1186/1471-2164-9-S1-S25

Zhang, D., Gao, F., Jakovlic, I., Zou, H., Zhang, J., Li, W. X., et al. (2020). PhyloSuite: An integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. *Mol. Ecol. Resour* 20 (1), 348–355. doi: 10.1111/1755-0998.13096

Zhang, Y., Du, L., Liu, A., Chen, J., Wu, L., Hu, W., et al. (2016). The complete chloroplast genome sequences of five epimedium species: Lights into phylogenetic and taxonomic analyses. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.00306

Zhang, Y. J., Ma, P. F., and Li, D. Z. (2011). High-throughput sequencing of six bamboo chloroplast genomes: phylogenetic implications for temperate woody bamboos (Poaceae: Bambusoideae). *PloS One* 6 (5), e20596. doi: 10.1371/journal.pone.0020596

Zhao, J., Yue, X. L., He, Z. R., and Zhou, X. M. (2021). The complete chloroplast genome of endangered species aristolochia delavayi franch. (Aristolochiaceae) in southwestern China. *Mitochondrial DNA B Resour* 6 (8), 2339–2341. doi: 10.1080/23802359.2021.1931506

Zhou, J., Chen, X., Cui, Y., Sun, W., Li, Y., Wang, Y., et al. (2017). Molecular structure and phylogenetic analyses of complete chloroplast genomes of two aristolochia medicinal species. *Int. J. Mol. Sci.* 18 (9), 1–15. doi: 10.3390/ijms18091839

Zhu, X.-X., Li, X.-Q., Liao, S., Du, C., Wang, Y., Wang, Z.-H., et al. (2019a). Reinstatement of isotrema, a new generic delimitation of aristolochia subgen. siphisia (Aristolochiaceae). *Phytotaxa* 401 (1), 1–23. doi: 10.11646/phytotaxa.401.1.1

Zhu, X., Wang, J., Liao, S., and Ma, J. (2019b). Synopsis of aristolochia l. and isotrema raf. (Aristolochiaceae) in China. *Biodiversity Sci.* 27 (10), 1143–1146.

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 22 March 2023
DOI 10.3389/fpls.2023.1160446

OPEN ACCESS

EDITED BY
Weijun Kong,
Capital Medical University, China

REVIEWED BY
Yunzhu Chen,
Hunan Academy of Forestry, China
Wenfeng Chen,
Fuzhou University, China
Mingying Zhang,
Shaanxi University of Chinese Medicine, China

*CORRESPONDENCE
Jinxin Liu
liujx_23@163.com
Shunxing Guo
sxguo1986@163.com

SPECIALTY SECTION
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

RECEIVED 07 February 2023
ACCEPTED 08 March 2023
PUBLISHED 22 March 2023

CITATION
Zhou L, Chen T, Qiu X, Liu J and Guo S (2023) Evolutionary differences in gene loss and pseudogenization among mycoheterotrophic orchids in the tribe Vanilleae (subfamily Vanilloideae). *Front. Plant Sci.* 14:1160446. doi: 10.3389/fpls.2023.1160446

Evolutionary differences in gene loss and pseudogenization among mycoheterotrophic orchids in the tribe Vanilleae (subfamily Vanilloideae)

Lisi Zhou, Tongyao Chen, Xiandan Qiu, Jinxin Liu* and Shunxing Guo*

Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

Introduction: *Galeola lindleyana* is a mycoheterotrophic orchid belonging to the tribe Vanilleae within the subfamily Vanilloideae.

Methods: In this study, the *G. lindleyana* plastome was assembled and annotated, and compared with other Vanilleae orchids, revealing the evolutionary variations between the photoautotrophic and mycoheterotrophic plastomes.

Results: The *G. lindleyana* plastome was found to include 32 protein-coding genes, 16 tRNA genes and four ribosomal RNA genes, including 11 pseudogenes. Almost all of the genes encoding photosynthesis have been lost physically or functionally, with the exception of six genes encoding ATP synthase and *psaJ* in photosystem I. The length of the *G. lindleyana* plastome has decreased to 100,749 bp, while still retaining its typical quadripartite structure. Compared with the photoautotrophic Vanilloideae plastomes, the inverted repeat (IR) regions and the large single copy (LSC) region of the mycoheterotrophic orchid's plastome have contracted, while the small single copy (SSC) region has expanded significantly. Moreover, the difference in length between the two *ndh*B genes was found to be 682 bp, with one of them spanning the IRb/SSC boundary. The Vanilloideae plastomes were varied in their structural organization, gene arrangement, and gene content. Even the *Cyrtosia septentrionalis* plastome which was found to be closest in length to the *G. lindleyana* plastome, differed in terms of its gene arrangement and gene content. In the LSC region, the *psb*A, *psb*K, *atp*A and *psa*B retained in the *G. lindleyana* plastome were missing in the *C. septentrionalis* plastome, while, the *mat*K, *rps*16, and *atp*F were incomplete in the *C. septentrionalis* plastome, yet still complete in that of the *G. lindleyana*. Lastly, compared with the *G. lindleyana* plastome, a 15 kb region located in the SSC area between *ndh*B-*rrn*16S was found to be inverted in the *C. septentrionalis* plastome. These changes in gene content, gene arrangment and gene structure shed light on the polyphyletic evolution of photoautotrophic orchid plastomes to mycoheterotrophic orchid plastomes.

Discussion: Thus, this study's decoding of the mycoheterotrophic *G. lindleyana* plastome provides valuable resource data for future research and conservation of endangered orchids.

KEYWORDS

mycoheterotrophic, Vanilloideae, plastome, gene loss, rearrangement, IR/SC boundaries

Introduction

The Orchidaceae, one of the two largest families of flowering plants, are unique in features such as their labellum, stamen, pollen block and dust-like seeds (Chase et al., 2015). They include approximately 28,000 species in 736 genera (Christenhusz and Byng, 2016), and are distributed throughout the world. Molecular phylogenetic research has revealed the Orchidaceae to be a monophyletic group, divided into five subfamilies, namely Apostasioideae, Cypripedioideae, Vanilloideae, Orchidoideae and Epidendroideae (Chase et al., 2015; Li et al., 2016). Among these, the Vanilloideae, Orchidoideae and Epidendroideae subfamilies include both photoautotrophic and mycoheterotrophic orchids (Merckx, 2013), with 232 mycoheterotrophic species in 43 genera. This feature makes them ideal material for research related to trophic transformation and gene structure change. Molecular evidence has shown that the rate of evolution in the Vanilloideae subfamily has variously accelerated and slowed down (Wang et al., 2018), and it is the first independent clade to include mycoheterotrophic species after the Apostasioideae subfamily (Kim et al., 2020). Therefore, further in-depth investigation of the Vanilloideae subfamily plastomes can provide more abundant evidence for the analysis of the genetic evolution mechanism, as well as establish a molecular basis of trophic-type changes in orchids in their early evolutionary stage.

Photosynthesis is known to be the primary means by which plants obtain their nutrients. However, limited by factors such as challenging living conditions or oxidative stress, plants have also adaptively evolved alternative survival strategies such as mycoheterotrophy and parasitism (Leake, 1994; Bidartondo, 2005; Merckx and Freudenstein, 2010; Westwood et al., 2010; Delannoy et al., 2011; Merckx, 2013; Wicke et al., 2016). While such changes in the survival strategies and vegetative organs of plants occurred, research has shown that traces of gene variation and selection remained in their plastomes (Wicke et al., 2011; Bromham et al., 2013; Petersen et al., 2015; Wicke et al., 2016). The transition from autotrophic to parasitic was accompanied by the loss of photosynthesis and housekeeping plastid genes (Wolfe et al., 1992; Funk et al., 2007), as well as the function, size (Lohan and Wolfe, 1998) and transcriptional association with essential genes (Wicke et al., 2013), all of which are linked to the retention or lost of the gene in the plastids of non-photosynthetic plants. Therefore, based on data published regarding the evolutionary transition from phototrophic to parasitic plastomes, an overall trend prediction for progressive gene loss has been proposed (Barrett and Davis, 2012; Barrett et al., 2014a; Bellot and Renner, 2016; Naumann et al., 2016; Wicke et al., 2016). In this hypothesis, the plastid genome reduction process is divided into the following four stages: 1) The loss of the NADH dehydrogenase-like (NDH-1) complex, generally regarded as the earliest loss of plastid-encoded genes; 2) This stage was followed by pseudogenization and the loss of photosynthetic genes, the deprivation of photosynthetic function and the removal of selection pressure to retain photosynthetic plastid genes; 3) Subsequently, the loss of genes for the plastid-encoded subunits of RNA polymerase and photosynthetic enzymes with minor functions (Rubisco and ATP synthase), the relative timing of which had an asynchronous and comparatively wide window; and 4) The delayed loss of the five core non-bioenergetic genes (especially *trn*E and *acc*D, which encode glutamyl tRNA and acetyl-CoA carboxylase subunits, respectively), observed only in fully parasitic plastomes with large-scale gene loss. The range of changes of mycoheterotrophic plastomes is similar to that of parasitic species (Barrett et al., 2014b; Lam et al., 2015; Lam et al., 2016; Kim et al., 2019; Kim et al., 2020), however, in the evolution of plastid genomes in mycoheterotrophic orchids belonging to the same tribe, gene variation and loss have been specific rather than convergent. Feng et al. (2016), for example, found that the transitions to a fully mycoheterotrophic lifestyle evolved independently at least three times during the evolution of the tribe Neottieae. Despite the general trend of plastid degradation was similar during the transition from autotrophic to mycoheterotrophic and from autotrophic to parasitic, highly lineage-specific plastome degeneration still occurred.

The Vanilloideae subfamily includes 14 genera and 245 species (Chase et al., 2015) with various lifestyles, including epiphytic and terrestrial. Only nine orchid plastomes have thus far been reported, distributed among four genera and two tribes (Cameron, 2009; Lin et al., 2015; Amiryousefi et al., 2017; Niu et al., 2017; Kim et al., 2019; Kim et al., 2020). Of the 14 Vanilloideae subfamily heterotrophs in five genera, only three species in two genera have been reported. Due to the lack of decoded mycoheterotrophic Vanilloideae plastomes, the study of genetic variation in plastomes during the lifestyle transition from photoautotrophic to mycoheterotrophic has thus far been insufficient, and the diversity of evolutionary models has not been verified. Thus, this study sequenced, assembled, and annotated the mycoheterotrophic *G. lindleyana* plastome, subsequently comparing it with the plastomes of previously researched photosynthetic and heterotrophic orchids

in the same subfamily, in order to explore the genetic variations occurring in plastomes that transitioned from photoautotrophic to mycoheterotrophic.

Material and methods

Plant materials, DNA extraction and high-throughput sequencing

Since the *G. lindleyana* plant does not have typical leaf organs, its root was used to obtain its chloroplast genome sequence. The root of *G. lindleyana* (Figure 1) was collected from Yuexi in Anhui Province, China (30.84°N; 116.34°E) at an altitude of 1130 m on October 2, 2020. This plant was identified morphologically, and its voucher specimen deposited in the Institute of Medicinal Plant Development, Chinese Academy of Medicinal Sciences.

Fresh root samples were ground into fine powder with liquid nitrogen in a mortar, and then used to extract the genomic DNA using a Plant Genomic DNA Extraction Kit (Tiangen Biotech (Beijing) Co., Ltd., China). The DNA concentration and the ratios of A260/A280 and A260/A230 were measured using a Thermo Scientific NanoDrop 2000 ultra-micro spectrophotometer (Thermo Fisher Scientific Inc., MA, USA). Following the construction of a 270 bp PCR-free library, the whole genomic DNA of *G. lindleyana* was sequenced using the Illumina NovaSeq 6000 platform *via* shotgun sequencing.

Plastome assembly and annotation

The resulting raw high-throughput sequencing reads were trimmed and filtered using Trimmomatic v0.38 (Bolger et al., 2014). The complete *G. lindleyana* chloroplast genome was assembled using the GetOrganelle(v1.7.7.0) toolkit (Jin et al., 2020), and four junction regions between IRs and LSC/SSC were subsequently confirmed *via* polymerase chain reaction (PCR) amplifications and Sanger sequencing using the primers listed in Table S1. The genome sequence was automatically annotated using the CPGAVAS2 integrated web server (Shi et al., 2019), and then manually edited using the Apollo editor according to separate BLASTN results comparing the CDS and protein sequences of 238 previously published plastid genomes in the Orchidaceae family.

Genome comparison, divergence and phylogenetic analysis

Genome features, such as size, LSC region, SSR region, IR region, GC content and unique genes, were analyzed or counted using a local Python script. SSRs were predicted *via* MISA(http://pgrc.ipk-gatersleben.de/misa/) microsatellite finder (Beier et al., 2017);, and the tandem repeat sequences were found using a tandem repeats finder (TRF) (Benson, 1999). SC-IR junction regions were described *via* IR Scope among six Vanilleae species (Amiryousefi et al., 2018), while chloroplast genome comparison was completed using mVISTA (http://genome.lbl.gov/vista/mvista/submit.shtml) with the annotation of four Vanilleae species as references (Frazer et al., 2004). Pairing sequence alignments of the cp genomes were performed using Mummer v3.23 with the six Vanilleae species and five mycoheterotrophic orchids (Kurtz et al., 2004). Synonymous codon usage and RSCU were analyzed using the CodonW v1.4.2 (Sharp and Li, 1987). The plastome sequence gene data of 59 Orchidaceae species were also used to construct a maximum likelihood (ML) phylogenetic tree, using five species (*Allium cepa, Eustrephus latifolius, Iris gatesii, Iris sanguinea* and *Fritillaria hupehensis*) as out groups. A total of 79 CDS genes were extracted and aligned using MAFFT software v7.515 (Katoh et al., 2002), and these alignments were subsequently concatenated into a single length of 71, 597 bp. The missing data of CDS genes were treated as insertions/deletions, and filled the alignments with dashs. Concatenated alignments were then used to perform the JModelTest. Finally, an ML tree was constructed *via* RAxML v8.2.12 (Stamatakis, 2014) with a GTR+G+I model with 1000 bootstrap replicates.

Results

Features of *G. lindleyana* cpDNA

A total of 43 million pair-end reads were produced with 5.81 Gb of clean data. Data from all of the reads were deposited in the NCBI Genbank under accession number MW528436. The complete plastome was found to be 100,749 bp (Figure 2), and displayed a typical quadripartite structure, including a pair of inverted repeat regions (IR; 11,607 bp) separated by large single copy (LSC; 59,493 bp) and small single copy (SSC; 18,042 bp) regions, covering 11.5%, 59.1% and 17.9% in the plastome, respectively (Table 1). The total GC content in the whole *G. lindleyana* plastome was 34.36%, with the LSC region containing the lowest amount of GC contents (31.24%) compared to those of the SSC (38.91%) and IR (38.80%) regions.

A total number of 63 genes were encoded in the plastome, including 32 protein coding genes, 16 tRNA, and 4 rRNA genes, including 11 pseudogenes (Table 2). The protein-coding genes included 12 genes encoding small subunits of ribosome (*rps*2, 3, 4, 7, 8, 11, 12, 14, 15, 16, 18 and 19), eight genes encoding large subunits of ribosome (*rpl*2, 14, 16, 20, 22, 23 and 33), six genes encoding ATP synthase (*atp*A, B, E, F, H, and I), two genes encoding conserved open reading frames (*ycf*1 and 2), only one gene related to photosystem I (*psaJ*), and four genes encoding subunits of acetyl-CoA-carboxylase (*acc*D), protease (*clp*P), translational initiation factor (*inf*A) and maturase (*mat*K), respectively. Among these unique genes, four (*atp*F, *rpl*2, *rps*16 and *trn*L) were found to have one intron each, while two genes (*clp*P and *rps*12) comprised two introns each. Almost all of the genes encoding photosynthesis had undergone pseudogenization, with the exception of six genes encoding ATP synthase and *psaJ* in photosystem I, which were still intact.

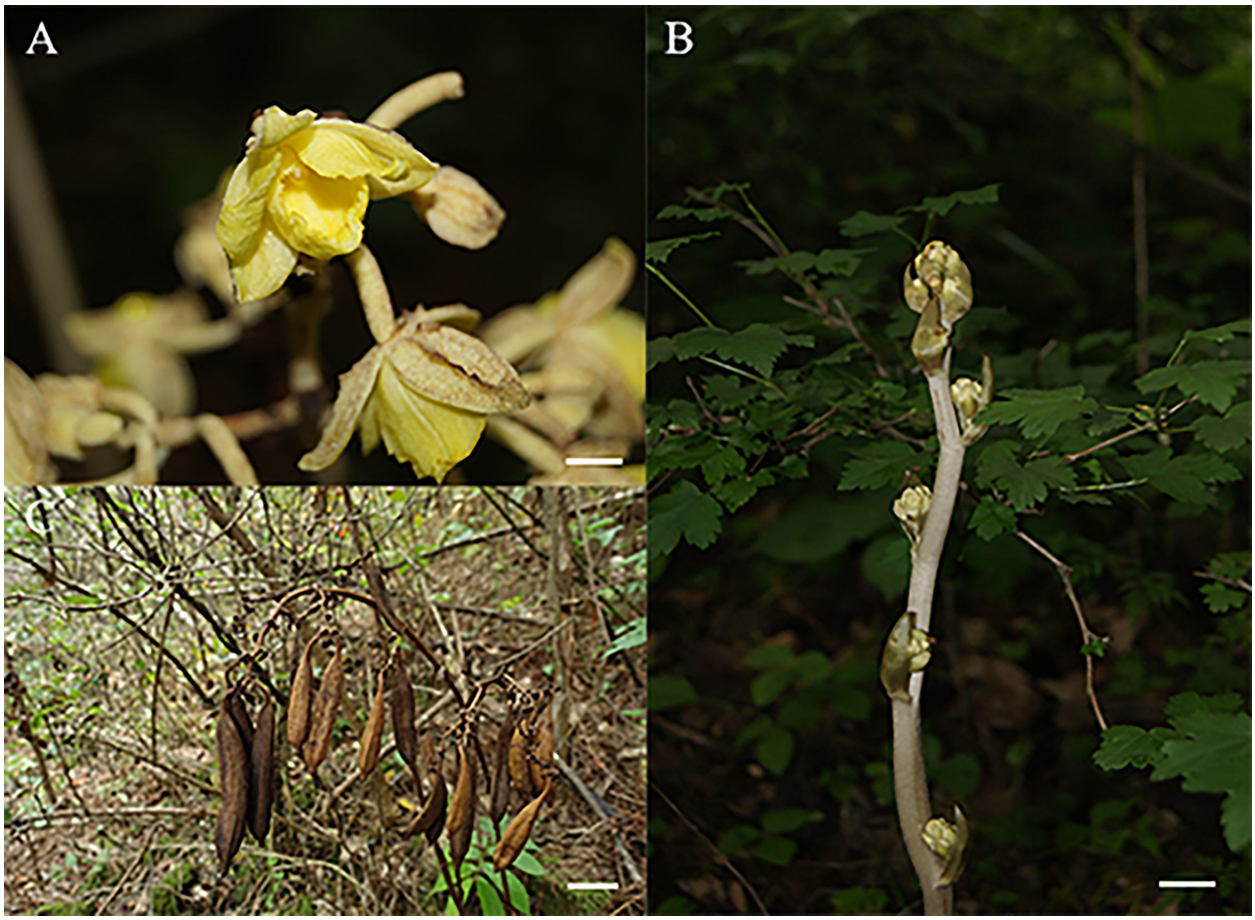


FIGURE 1
Morphological characteristics of *G. lindleyana:* **(A)** Flower, scare bar = 1cm; **(B)** stem, scare bar = 1cm; **(C)** capsule, scare bar = 5cm.

In this plastome the most frequently used codon was AAA (n=1374) followed by AAT (n=1032), encoding lysine and asparagine, respectively. The least frequently used codon was stop codon TGA (n=31).

Codon usage

Codon usage patterns of the coding sequences for *G. lindleyana* were calculated based on the relative synonymous codon usage (RSCU) value. All protein-coding genes in *G. lindleyana* plastome were encoded by 9749 codons (Figure S1). A total of 61 codons encoded 20 amino acids, with three stop codons. There were 32 preferred and 32 non-preferred codon usages. Among them, the most abundant amino acid was leucine, with 962 codons (9.87% of total), followed by isoleucine with 838 codons (8.60% of total) (Figure S1A), while stop codons were the fewest, at just 31 (0.32% of total). Almost all amino acids had more than one synonymous codon, with the exceptions of tryptophan and methionine. Arginine, serine and leucine had the most synonymous codon usages.

The RSCU value was used to evaluate the synonymous codon bias, and codons with greater RSCU values were preferred in each case (Figure S1B). The most preferred codon was found to be AGA encoding amino acid arginine (Arg), with 1.99 RSCU, followed by UUA encoding leucine (Leu) and UCU encoding serine (Ser) with 1.79 and 1.73 RSCU, respectively. By contrast, the lowest frequency codon was found to be CAC encoding histidine (His) with 0.32 RSCU, followed by CGC encoding arginine (Arg) with 0.37 RSCU. With the exceptions of the leucine encoded by UUG and serine encoded by UCC, amino acid codons (RSCU > 1) in the *G. lindleyana* plastome preferentially showed A- or U-endings, and non-preferred codon usages ended with base C or G, corresponding to the previously mentioned results that were calculated based on protein-coding sequences.

Simple sequence and tandem repeat analyses

Simple sequence repeats (SSRs, also known as microsatellites) and tandem repeats (TRs) in the *G. lindleyana* plastome were surveyed in this study. A total of 41 SSRs were detected, following strict performance parameters (unit_size/min_repeats): 1/10 (mononucleotides ≥ 10 nt), 2/6 (dinucleotides ≥ 6 repeats), 3/5, 4/5, 5/5 and 6/5 (Table S1). Analysis included 29 mononucleotide repeats, nine dinucleotide repeats and two

10.3389/fpls.2023.1160446

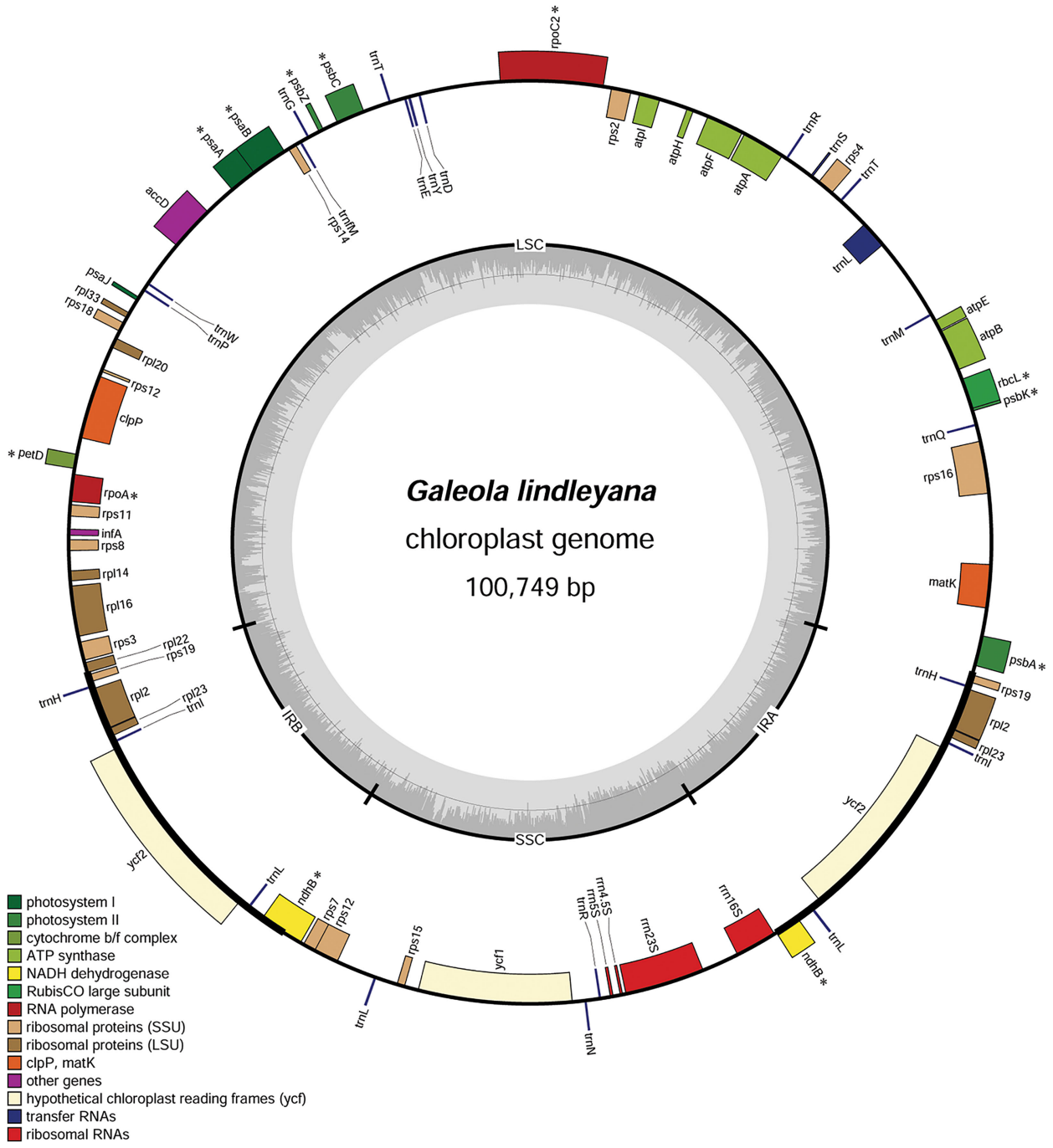


FIGURE 2
Chloroplast genome map of *G. lindleyana*. Genes inside the circle are transcribed clockwise, whereas those outside are transcribed counterclockwise. The darker gray in the inner circle corresponds to the GC content, while the lighter gray corresponds to the AT content. Pseudogenes are marked with an asterisk (*).

trinucleotide repeats, with only one hexa-nucleotide SSR identified. No tetra- or penta-nucleotide SSRs were found. The majority of SSRs were mononucleotides repeats (70.7%) in which base T and A were the primary elements with a proportion of 68.3%, only one C motif and no G motifs. The T-repeat unit was found to be most abundant in this study, with a total of 16, while the hexa-nucleotide SSR was repeated most frequently at 23 times in total.

There were 116 copy number variations (CNVs) of TR units (Table S1) in the *G. lindleyana* plastome ranging in length from 7 to 63 bp, with those of 16 and 18 bp most abundant (12), followed by those of 12 bp (10), and then those of 13 bp (7) and 24 bp (7). Most of the 116 CNVs were found to be present in intergenic regions, and only nine were present in the genic regions of the *acc*D, *ycf*1 and *ycf*2 genes, and the *psb*C pseudogene. The TR units

10.3389/fpls.2023.1160446

TABLE 1 Chloroplast genome features of *G. lindleyana* and its closely related species in the Vanilleae tribe.

Taxon Name	Life style	Nutritional mode	Genome Size/bp	LSC Length/ bp	SSC Length/ bp	IR Length/ bp	GC Content/ %	No. of coding genes	No. of tRNAs
Galeola lindleyana	Terrestrial	Mycoheterotrophic	100749	59493	18042	11607	34.4	32	16
Cyrtosia septentrionalis	Terrestrial	mycoheterotrophic	96859	58085	17946	10414	34.8	41	25
Lecanorchis japonica	Terrestrial	mycoheterotrophic	70498	28197	14493	13904	30.4	25	7
Lecanorchis kiusiana	Terrestrial	mycoheterotrophic	74084	30824	14118	14571	30.0	25	8
Vanilla aphylla	Epiphyte	Photosynthetic	150165	87379	3354	29716	35.0	65	29
Vanilla madagascariensis	Epiphyte	Photosynthetic	151552	87490	1254	31404	34.6	71	29
Vanilla planifolia	Epiphyte	Photosynthetic	148011	86358	2037	29808	35.4	72	30
Vanilla pompona	Epiphyte	Photosynthetic	148009	86358	2037	29807	35.4	72	30

appeared more frequently in the LSC regions (81.0%) than in the SSC (12.9%) and IR (6.0%) regions. CNVs of various TR units related to indel polymorphism were also identified in the *G. lindleyana* plastome.

Junctions of inverted repeats and single copy regions

A comprehensive assessment of the four junctions (J_{LA}, J_{LB}, J_{SA} and J_{SB}) between the two IR regions and the two single copy regions (LSC and SSC) of six species in the subfamily Vanilloideae was also performed, and the results are presented in Figure 3. Evidence of substantial expansion and contraction in both the IR regions and the two single copy regions were detected, with the IR regions ranging from 32,683 bp in *Pogonia japonica* to 10,414 bp in *Cyrtosia septentrionalis*; the LSC region ranging from 87,447 bp in *P. japonica* to 28,197 bp in *Lecanorchis japonica*; and the SSC region ranging from 18,042 bp in *G. lindleyana* to 2,146 bp in *Vanilla aphylla*.

Two junctions were conserved between the LSC region and the two IR regions in the same genus, but were variated in different genera. The distance between the J_{LA} border and *mat*K gene was found to be the same in all *Lecanorchis* plastomes. Similarly, the *rpl*2 gene that crossed J_{LB} in the two *Lecanorchis* plastomes was located in the LSC region but expanded to 127 bp to reach the IRb region. However, the J_{LA} and J_{LB} borders were variable in the other four species: The J_{LB} borders of *P.*

TABLE 2 Gene composition in *G. lindleyana* chloroplast genome.

Gene functions	Gene set	Gene
Photosynthesis	Photosystem I	*psa*A, *psa*B, *psa*J
	Photosystem II	*psb*A, *psb*C, *psb*K, *psb*Z
	Cytochrome b/f complex	*pet*D
	ATP synthase	*atp*A, *atp*B, *atp*E, *atp*F*, *atp*H, *atp*I
	NADH-dehydrogenase	*ndh*B×2
	Rubisco	*rbc*L
Self replication	Large subunit of ribosome	*rpl*14, *rpl*16, *rpl*2×2*, *rpl*20, *rpl*22, *rpl*23×2, *rpl*33
	Small subunit of ribosome	*rps*11, *rps*12**, *rps*14, *rps*15, *rps*16*, *rps*18, *rps*19×2, *rps*2, *rps*3, *rps*4, *rps*7, *rps*8
	DNA dependent RNA polymerase	*rpo*A, *rpo*C2
Other gene	rRNA genes	*rrn*16S, *rrn*23S, *rrn*4.5S, *rrn*5S
	tRNA genes	*trn*D, *trn*E, *trn*G, *trn*H, *trn*I, *trn*L*, *trn*M, *trn*N, *trn*P, *trn*Q, *trn*R, *trn*S, *trn*T, *trn*W, *trn*Y, *trnf*M
	Subunit of Acetyl-CoA-carboxylase	*acc*D
	Protease	*clp*P**
	Translational initiation factor	*inf*A
	Maturase	*mat*K
Unkown	Conserved open reading frames	*ycf*1, *ycf*2×2

"×2" indicates that the number of repeat units is two; One or two asterisks following genes indicate one or two contained introns, respectively. Pseudogenes are marked with an underscore.

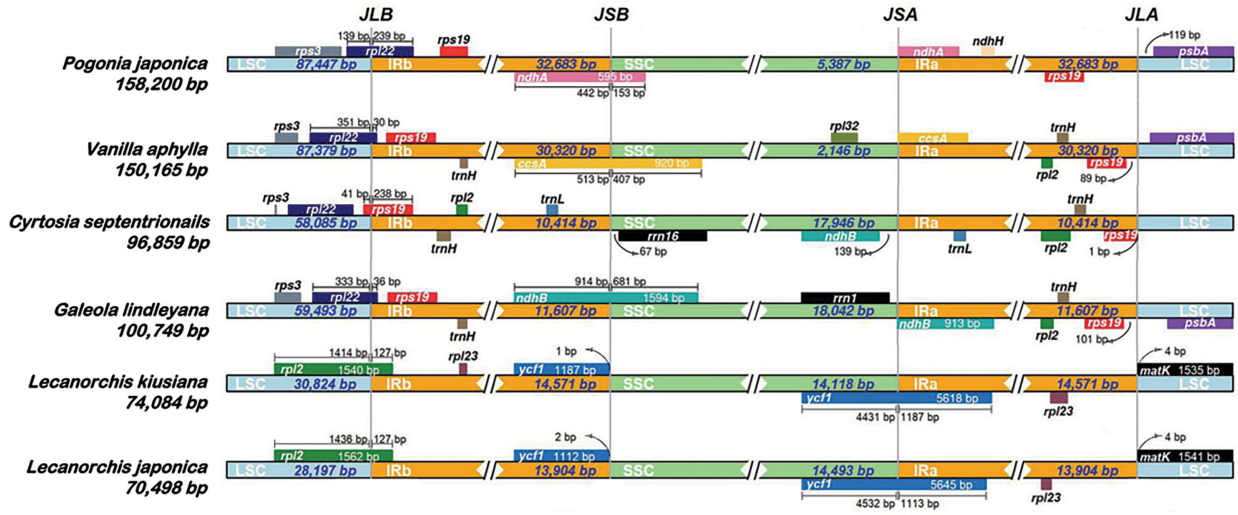


FIGURE 3
Comparisons of LSC, IRb, SSC and IRa border regions in six species of Vanilleae.

japonica, *V. aphylla* and *G. lindleyana* plastomes were within the *rpl*22, with some of the sequences present in the LSC region, showing expansions of 239 bp, 30 bp and 36 bp to reach the IRb region, respectively, while the *C. septentrionalis* plastome exhibited LSC contraction, leading to the entire *rpl*22 in the LSC region.

The IR contraction in the *C. septentrionalis* plastome brought about the crossing of *rps*19 over the J_{LB} border, with 41 bp located in the LSC region and the entire *rps*19 in the IRb region of the *P. japonica*, *V. aphylla* and *G. lindleyana* plastomes. In addition, J_{LA} was found between *rps*19 and *psb*A in *P. japonica*, *V. aphylla* and *G. lindleyana* plastomes, although the *psb*A was not present at the J_{LA} border in *C. septentrionalis* plastome. The distances between *rps*19 and J_{LA} were found to be 1 bp in *C. septentrionalis* plastome, 89 bp in *V. aphylla* plastome and 101 bp in *G. lindleyana* plastome, respectively, all of which were longer than the distance of 119 bp between *psb*A and J_{LA} in *P. japonica* plastome.

In these six species belonging to the Vanilleae, more variations were found in the J_{SA} and J_{SB} than in the J_{LA} and J_{LB}. In *Lecanorchis kiusiana* and *L. japonica* plastomes, the distance between *ycf*1 and J_{SB} was 1 bp and 2 bp, respectively, and the entire *ycf*1 was in the IRb region. However, the *ycf*1 in these two *Lecanorchis* species crossed the J_{SA} border at 4431 bp and 4532 bp located in SSC region and 1187 bp and 1113 bp in the IRa region in *L. kiusiana* and *L. japonica* plastomes, respectively. The genes around the J_{SA} and J_{SB}, and its location were varied in the other four species. In *C. septentrionalis* plastome, *rrn*16 and *ndh*B were located entirely in the SSC region, at 67 bp from the J_{SB} border and 139 bp from the J_{SA} border, respectively.On the J_{SB} border, the *ndh*A in *P. japonica*, *ccs*A in *V. aphylla* and *ndh*B in *G. lindleyana* crossed the IRb/SSC boundary, with 442 bp, 513 bp and 914 bp of each gene located in the IRb regions, respectively, and 153 bp, 407 bp and 1594 bp within the SSC regions, respectively. Moreover, the *ndh*A in *P. japonica*, *ccs*A in *V. aphylla* and *ndh*B in *G. lindleyana* were all at the beginning of the IRa regions, near the J_{SA} borders.

Comparative analysis and sequence divergence analyses

The differences in the four mycotrophic Vanilleae plastomes were evaluated using the phototrophic *V. aphylla* plastome as a reference, and the results are presented in Figure 4. Compared with the phototrophic species, the deletion of gene sequences in mycotrophic species was found to be significant. In *L. kiusiana* and *L. japonica*, all functional genes involved in photosynthesis have been lost. In *C. septentrionalis* and *G. lindleyana*, most genes of the photosystem were lost, while others, such as *psa*A, *psa*B and *psa*C, have been retained. Among these five species, the length of *G. lindleyana* and *C. septentrionalis* plastomes were found to be similar, but still showed significant differences in their functional gene loss and pseudo-genes. In the LSC area, the *psb*A, *psb*K, *atp*A and *psa*B retained by *G. lindleyana* were found to be missing in *C. septentrionalis*; while the *mat*K, *rps*16 and *atp*F present as pseudogenes in *C. septentrionalis* remain complete in *G. lindleyana*.

Dynamic chloroplast genome structures of Vanilleae and three typical mycotrophic orchids

Gene block analysis, by which gene rearrangements are identified, was carried out using Mauve software among six Vanilleae orchids and four typical mycotrophic orchids (Figure 5). In the Vanilleae, rearrangement and inversion of genes appeared frequently, and even the phototrophic orchid plastomes were not fully colinear in all regions. In *P. japonica* and *V. aphylla*, approximately 37 kb of inversion occurred in the LSC area. However, compared with the other four mycotrophic orchid plastomes, gene loss and rearrangements of these two phototrophic orchid plastomes were limited. Among the

10.3389/fpls.2023.1160446

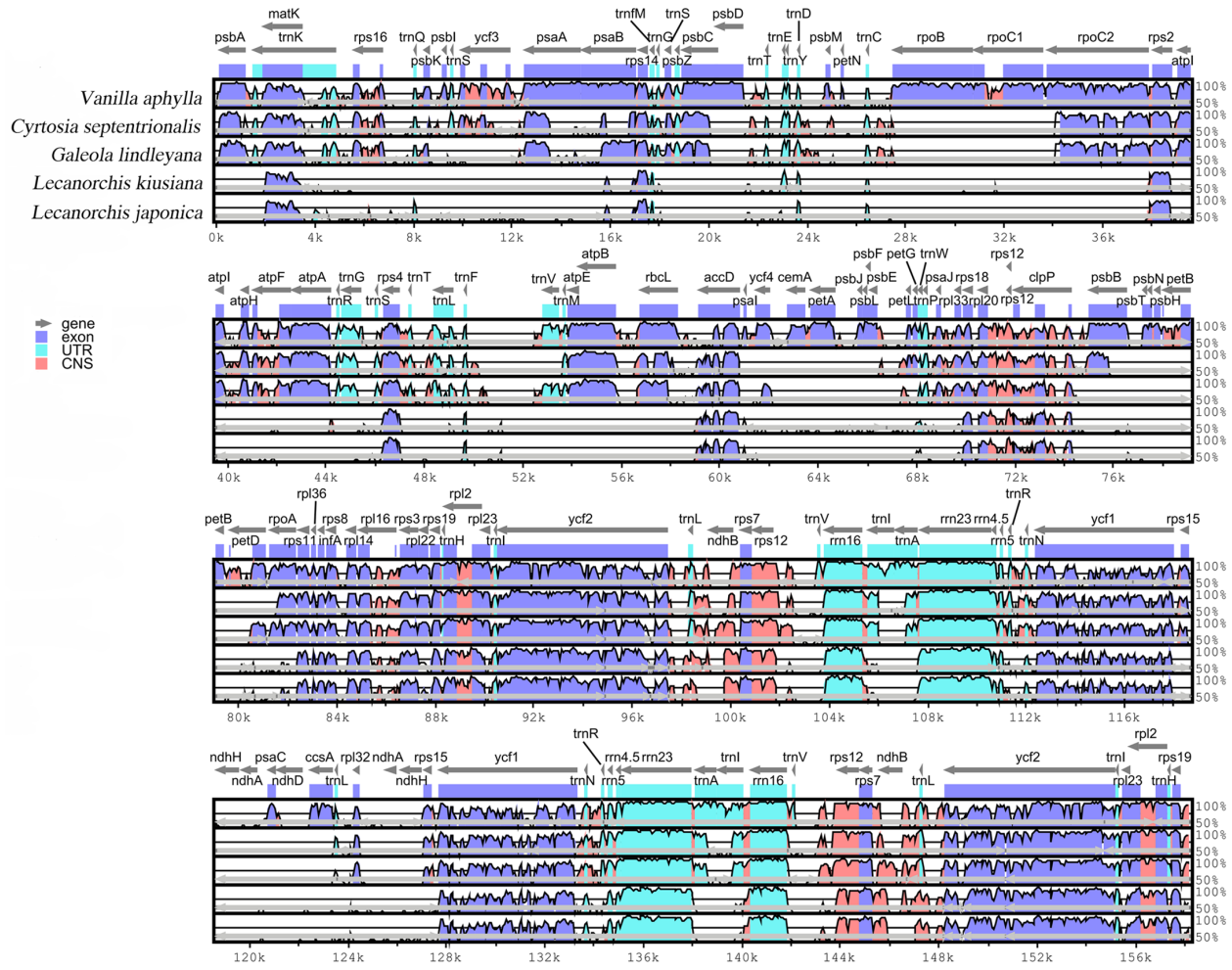


FIGURE 4
Global alignment of five chloroplast genomes of Vanilleae using mVISTA. Y-axis indicates the range of identity (50%-100%). Alignment was performed using *V. aphylla* as reference.

Vanilleae, the species with the chloroplast genome most similar to that of *G. lindleyana* was *C. septentrionalis*, however, with a 15 kb inverted in the SSC area, between *ndh*B and *rrn*16S. Compared with *G. lindleyana*, the photosynthesis genes in the LSC regions of the *L. kiusiana* and *L. japonica* chloroplasts were absent, and the *rpl2-rps*19 in the IR area was inverted to the LSC area.

Phylogenetic relationships

In phylogenetic analysis, the Orchidaceae were divided into five subfamilies using the classification of Orchidaceae proposed by Chase et al. (2015). A total of 59 orchid plastomes from three of these subfamilies were employed to infer the phylogenetic relationship, using the maximum likelihood (ML) methods (Figure 6). Support values were consistently very high, except for the branch leading to the tribes Epidendreae, Neottieae and Cranichideae. Eight species belonging to Vanilleae formed a monophyletic group with high support. In the tribe Vanilleae, four mycoheterotrophic species were well distinguished from the photoautotrophic species and, within them, genus *Lecanorchis* was resolved as sister to the branch including *G. lindleyana* and *C. septentrionalis*. The respective branch support values for the branch including *G. lindley*ana, *C. septentrionalis* and genus *Lecanorchis* were lower than those for the branch including *G. lindley*ana and *C. septentrionalis*. The mycoheterotrophic plastomes belonging to the tribe Vanilleae are divided into two main clades, one of which is represented by genus *Lecanorchis*, while the other group contains two genera, namely *Galeola* and *Cyrtosia*. Among these three genera, the *Galeola* and *Cyrtosia* plastomes were found to have lost a similar length and number of genes, with a greater loss than that of *Lecanorchis*. Since only these four mycoheterotrophic chloroplast genomes have thus far been reported in Vanilleae, more sample data is needed to confirm any differences between the chloroplast genome loss strategies of the two genera species (*Galeola* and *Cyrtosia*).

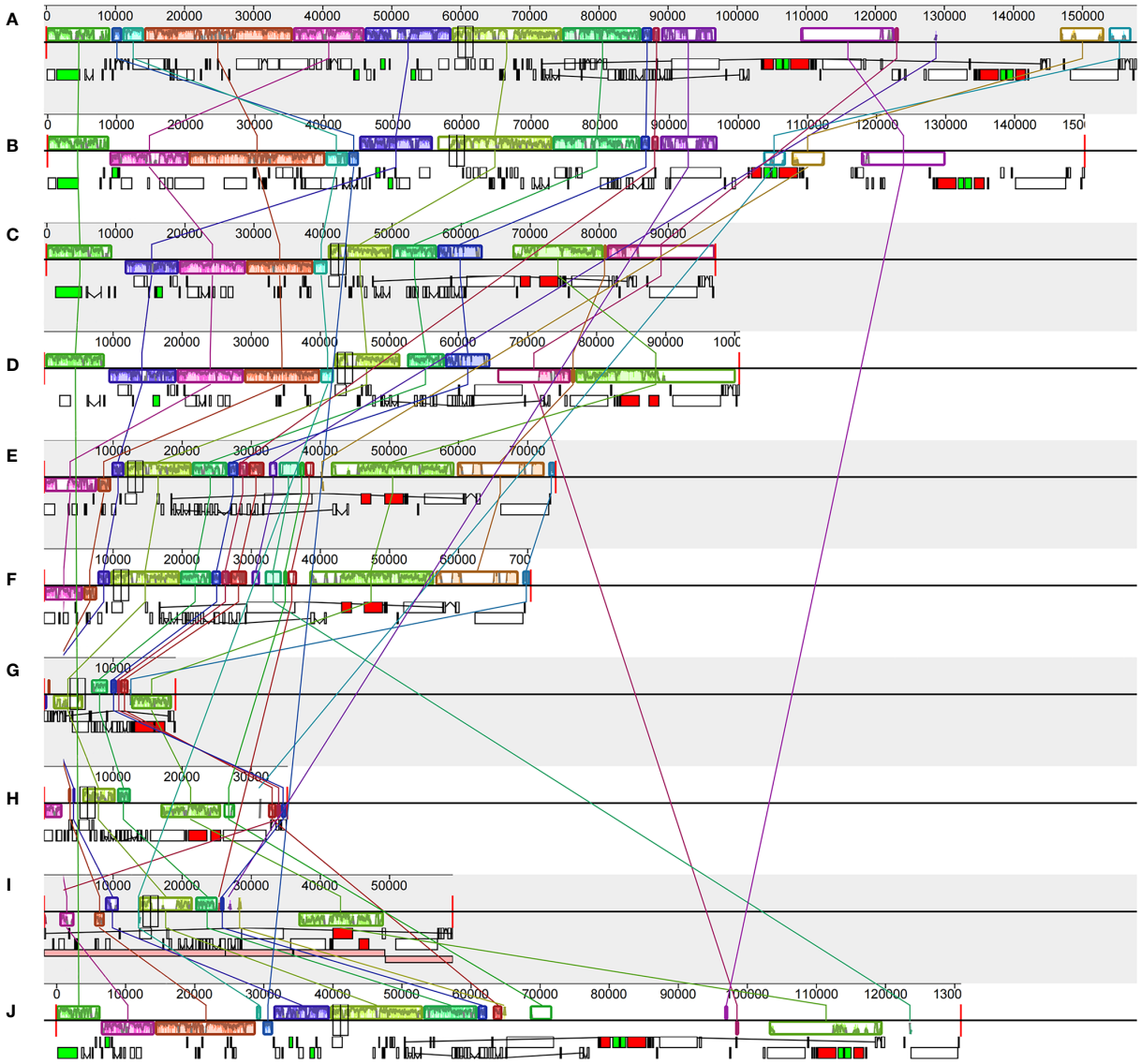


FIGURE 5
Comparisons of the complete chloroplast genome of six Vanilloideae species **(A)**. *P. japonica*; **(B)** *V. aphylla*; **(C)** *C. septentrionalis*; **(D)** *G. lindleyana*; **(E)** *L. kiusiana*; **(F)** *L. japonica*) and four typical mycoheterotrophic orchids **(G)**. *Epipogium roseum*; **(H)** *Gastrodia elata*; **(I)** *Rhizanthella gardneri*; **(J)** *Chamaegastrodia shikokiana*).

Discussion

Rearrangements in plastomes

Although chloroplast genome structures are generally highly conserved, some do undergo changes during the long-term evolutionary process, subsequently appearing as either rearrangements or inversions. Rearrangements and inversions were detected in all chloroplast genomes of both the autotrophic and heterotrophic orchids in the Vanilloideae subfamily. In the *Pogonia* and *Vanilla* genera, rearrangement and inversion were seen in the LSC, IR and SSC regions, while these changes were seen only in the IR and SSC regions in the *Cyrtosia* and *Galeola* genera, and were confined to the LSC region in *Lecanorchis* genus. Rearrangements of the IR and SSC regions occur frequently and are thought to be the product of species evolution (Jansen and Palmer, 1987; Doyle et al., 1996; Chumley et al., 2006; Lee et al., 2007; Jansen et al., 2008; Ravi et al., 2008; Weng et al., 2014; Yan et al., 2017; Frailey et al., 2018; Roma et al., 2018). In *Pelargonium hortorum*, for example, the IR region of the chloroplast genome has extensively expanded, increasing in length to 76 kb, which is three times that of most common plants (Chumley et al., 2006). By contrast, the IR regions in Fabaceae and Cupressaceae have contracted significantly and even, in some, been completely lost (Saski et al., 2005; Hirao et al., 2008). In addition,

10.3389/fpls.2023.1160446

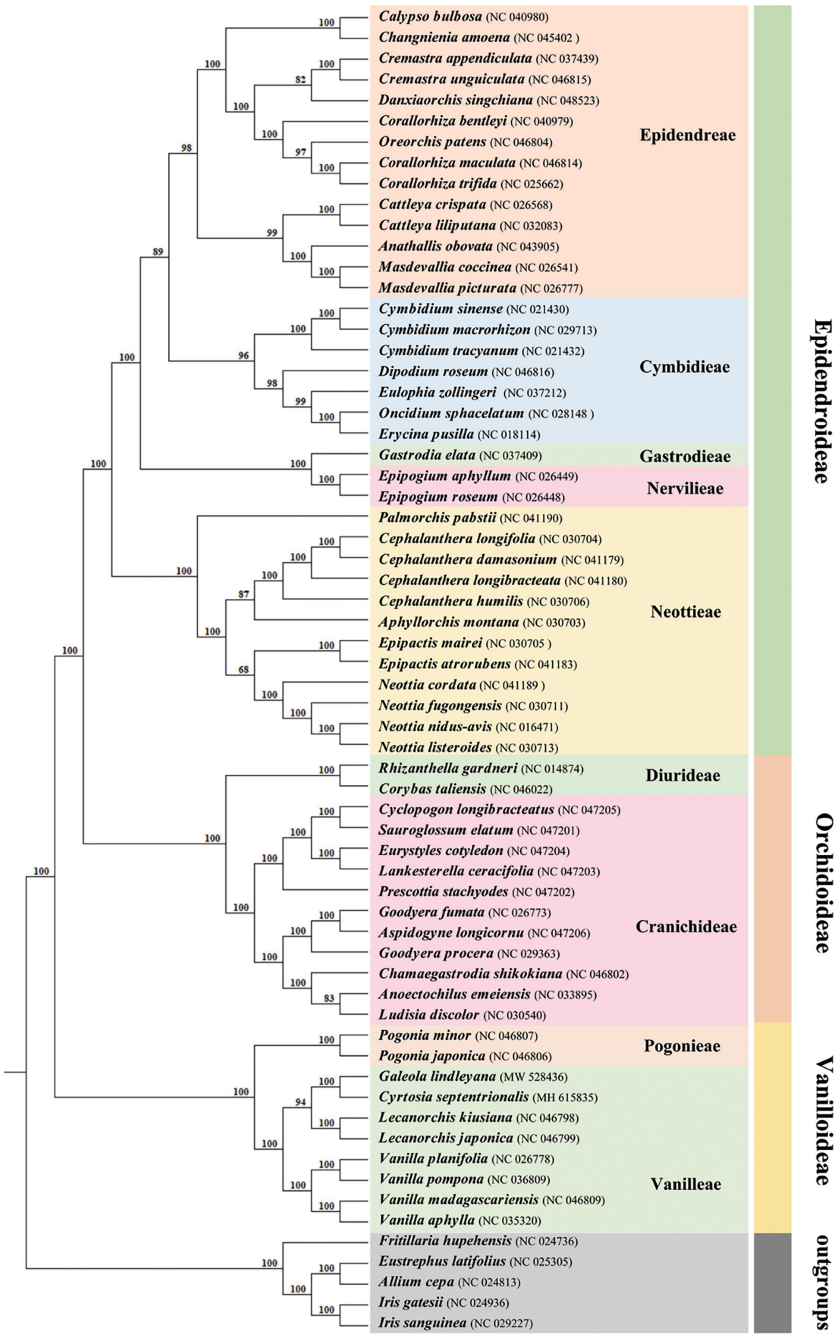


FIGURE 6
Phylogenetic tree for 59 Orchidaceae (with five non-Orchidaceae species as outgroups) using maximum likelihood (ML), based on the alignments of complete chloroplast genomes. Numbers at the nodes indicate bootstrap values from 1000 replicates.

the rearrangement of SSC regions has been detected in the chloroplast genomes of semi-parasitic species (Frailey et al., 2018), and the inversion of the SSC region, which appears to be a hallmark of the semi-parasitic group in the Orobanchaceae, has a chloroplast genome size similar to that of autotrophs. However, such rearrangement and inversion in all regions, as well as the initial contraction followed by expansion of the IR region, are rare, which suggests that the Vanilloideae subfamily may have been at the forefront of this evolutionary process. Changes in the nutrient types of the Vanilloideae subfamily, thus, closely related to the frequent occurrence of rearrangement and inversion.

Variations in the IR/SC boundary

All chloroplast genomes have a typical four-part structure in the Vanilloideae subfamily, as do mycoheterotrophic orchids. However, differences have been found within the mycoheterotrophic orchids

in the Orchidaceae subfamily, such as the loss of the IR region in *Gastrodia elata* plastome (Yuan et al., 2018; Kang et al., 2020; Park et al., 2020). The sizes and structures of chloroplast genomes in the Vanilloideae subfamily differ among the genera, and rearrangement and inversion occur in all four regions (Kim et al., 2015; Zeng et al., 2017). Among them, the longest chloroplast genome, at approximately 158,200 bp, is that of *P. japonica*, a member of the Pogonieae tribe, while the shortest, less than half the size at approximately 70,498 bp, is that of *L. japonica* belonging to the Vanilleae tribe. Such a dramatic difference in chloroplast genome length has also been found among the Neottieae tribe (Feng et al., 2016). However, some semi-heterotrophic species have also been found in the Orobanchaceae family, whose length of plastomes are similar to those of autotrophic species, suggesting that their plastomes are still in the early stages of their transition from autotroph to parasite (Wicke et al., 2013; Wicke and Naumann, 2018). The size of autotrophic orchid plastomes has been found to range from 147 kb to 151 kb, while those of mycoheterotrophic orchid plastomes fluctuate more widely, ranging from 70 kb to 100 kb. In mycoheterotrophic orchids, the plastomes of the genus *Lecanorchis* are smaller than 100 kb, while those of the genera *Cyrtosia* and *Galeola* are approximately 100 kb. Plastomes of the *Lecanorchis* orchid have not contracted significantly, possibly because the significant shrinkage of the LSC region is countered by the expansion of the SSC region. A similarly remarkable expansion of the SSC region to the IR region has also been reported in the Orobanchaceae family: Unexpectedly enlarged overall chloroplast genome lengths, attributable to the expansion of IR regions into SSC regions, have been noted in the semi-heterotrophic *Buchnera americana* as well as in three species belonging to the genus *Striga* in the same family (Frailey et al., 2018). Therefore, it is evident that the expansion of the SSC region in fully heterotrophic plants slowed down the pace of their plastome shrinkage, however, the reasons for this expansion still need to be explored. The expansion of SSC regions in the chloroplast genomes of semi-heterotrophic plants in the Orobanchaceae family and mycoheterotrophic orchids in the Vanilloideae subfamily is highly significant because it may indicate the evolution of key genes for energy metabolism in non-photosynthetic plants.

The IR/SC boundary and its nearby genes were analysed, among six Vanilloideae plastomes. and variations were found at all IR/SC boundaries. The shrinkage or expansion of the IR is known to lead to greater variability in the IR/SC boundary, and has been a recent hotspot in scientific research, especially in the Orchidaceae family (Wu et al., 2009; Wu et al., 2011; Sanderson et al., 2015). The LSC region has contracted in each of the six Vanilloideae plastomes, ranging from photosynthetic to mycoheterotrophic, with the LSC/IRB boundary gradually shrinking from *rpl*22 (*P. japonica*, *V. aphylla* and *G. lindleyana*) to *rps*19 (*C. septentrionalis*) and then to *rpl*2 (*L. kiusiana* and *L. japonica*). Moreover, the IR regions have contracted and then expanded from photosynthetic to mycoheterotrophic plastomes and the location of cross-boundary genes in the two boundaries was asymmetrical. The cross-boundary genes located in IRB/SSC and IRA/SSC boundaries changed gradually from *ndh*A (*P. japonica*) to *ccs*A (*V. aphylla*) to *ndh*B (*G. lindleyana*). However, the IR regions in *L. kiusiana* and *L. japonica* plastomes expanded,so that *ycf*1 regained its location in the IR regions. All Vanilloideae plastomes have genes across the IRB/SSC or IRA/SSC boundaries, with the exception of *C. septentrionalis*. Therefore, the cross-boundary genes are located in two IR regions and the SSC region simultaneously. In addition, the IRA/LSC boundary was found to have shrunk and then expanded. In *P. japonica*, *V. aphylla*, *G. lindleyana* and *C. septentrionalis* plastomes, the IRA region shrank gradually, and the *rps*19 moved to the IRA/LSC boundary gradually, while in *L. kiusiana* and *L. japonica* plastomes, the IRA region expanded until the *mat*K gene was located in the IRA/LSC boundary. These changes in the IR/SC boundaries are consistent with the two evolutionary routes in the Orchidaceae family (Yang et al., 2013; Luo et al., 2014): (1) The *ycf*1 gene expands continually into the IRA region, so the duplicated pseudogene *ψycf*1 fragment appearanced in the IRB region and overlapped with *ψndh*F. This evolutionary route was matched with changes in the heterotrophic orchid(*L. kiusiana* and *L. japonica*) plastomes. The SSC region expanded, causing the *ycf*1 to relocate in the IRA region; (2) The continuous movement of *ycf*1 to the SSC region resulted in the shorter length of the replicated *ycf*1 in the IRB region, which eventually moved completely into the SSC region, while the replicated *ycf*1 fragment disappeared in the IRB region. In this study, the transformation of plastomes from autotrophic (*P. japonica* and *V. aphylla*) to heterotrophic (*G. lindleyana* and *C. septentrionalis*) was found to be consistent with this hypothesis. The *ycf*1 was eventually located in the SSC region (*G. lindleyana* and *C. septentrionalis*). Changes in the spanning boundary genes were the major motivation for the contraction or expansion of IR regions (Goulding et al., 1996; Yang et al., 2010). Both of these hypotheses could well explain the changes at the IR/SC boundaries in the Orchidaceae family. Nevertheless, the evolutinonary origin and the chronological framework of cross-boundary gene changes were still largely unknown. Some researchers have shown that changes at the IR/SC boundaries could provide evidence for phylogenetic relationships (Gao et al., 2009; Wu et al., 2009; Wu et al., 2011; Yang et al., 2013; Downie and Jansen, 2015; Sanderson et al., 2015), however, in the Vanilloideae subfamily, the decoded plastomes were limited. In order to relate the changes at the IR/SC boundaries with phylogenetics, more Vanilloideae plastomes with various lifestyles should be collected and explored.

Genes loss in plastid genomes and diversity of mycoheterotrophs in Vanilloideae orchids

In this study, the number of genes in the plastomes of each species in the Vanilloideae subfamily was found to be different. Pseudogenes and gene loss were present in the chloroplast genomes of all species. Most of the *ndh* gene family have been lost fully or partially, such as the remaining residual fragments of *ndh*A, *ndh*B,

*ndh*D and *ndh*H. Rampant independent loss of the *ndh* gene family was a common feature across the Orchidaceae family (Kim et al., 2020) and, regardless of whether orchids were identified as mycoheterotrophic species or not, most of the *ndh* gene family were lost in the Vanilloideae subfamily orchid plastomes, with only the *ndh*B residual fragment still evident or a complete disappearance of the *ndh* gene (Kim et al., 2019; Kim et al., 2020). Assuming that the plastid genome degrades in a gradual manner, it may be also assumed the nonessential *ndh* gene pseudogenized before its fully physical deletion (Wicke et al., 2016; Wicke and Naumann, 2018). Pseudogenes could have been generated by premature codon termination and frameshift mutations (Balakirev and Ayala, 2003; Poliseno et al., 2010). The intron region of the *ndh*B gene in *G. lindleyana* plastome was found to have been shortened and was located at the IR/SC boundary, both favorable conditions for pseudogene generation. Moreover, the *ndh* gene family members were found to vary frequently and contained abundant repetitive sequences. Repeated sequences are considered to be one of the main reasons behind the rearrangement of the chloroplast genome (Yue et al., 2008). Therefore, the non-functionalization of the Vanilloideae subfamily chloroplast genome probably began with the pseudogenization and loss of the *ndh* gene family, as is consistent with the known evolutionary process of trophic changes (Wicke et al., 2013; Zeng et al., 2017). However, non-functionalization of the *ndh* gene and/or its fully physical loss was independent of trophic type in this species. Its polygenicity has been previously demonstrated in orchids (Barrett and Davis, 2012), carnivorous plants (Wicke et al., 2014) and even gymnosperms (Lin et al., 2010). In addition, Chang et al. (2006) found that *ndh*A, *ndh*F and *ndh*H in the *Phalaenopsis aphrodite* plastome had been transferred to its nuclear genome. Nonetheless, due to the lack of relevant nuclear genome data, it remains to be confirmed whether the *ndh* gene family has been transferred to its nuclear genome in the photosynthetic species of the Vanilloideae subfamily.

Currently, several models of chloroplast genome degeneration have been proposed to explain the physical or functional changes associated with the transition to a mycoheterotrophic lifestyle (Wicke et al., 2011; Bromham et al., 2013; Petersen et al., 2015; Wicke et al., 2016). In this study, photosynthetic plants were found to be in their initial stage of chloroplast genome degeneration, and the *ndh* gene appeared to be first pseudogenized and then lost entirely (Graham et al., 2017). During this period, reduced photosynthetic function begins with the loss of non-essential or stress-related genes (such as ndh genes) in half-heterotrophs. This is followed by pseudogenization and the loss of major photosynthesis-related genes (such as *pet*, *psa* and *psb* genes) and plastid-encoded polymerases. This was evident in this study, with both *G. lindleyana* and *C. septentrionalis* (Kim et al., 2019) found to be in this second stage of chloroplast genome degeneration. Among the photosynthesis-related genes, only *pet*N, *psa*J, *psb*M and *psb*Z remain in *C. septentrionalis*, while *G. lindleyana* no longer contains a complete gene of the photosynthesis-related genes and *pet*D, *psa*A, *psa*B, *psb*A, *psb*C, *psb*K and *psb*Z have all pseudogenized. The next stage of plastome degeneration ocurred in the edge of these shift from semi-mycoheterotrophic orchids to fully mycoheterotrophic orchids. Genes with extended or alternative functions, such as *atp* and *rbc*L, and non-essential housekeeping genes are lost after the transition to a non-photosynthetic or fully heterotrophic lifestyle, but prior to a plateau or slowing in the rate of gene loss. Herein, *L. kiusiana* and *L. japonica* (Kim et al., 2020) were found to be at this third stage of the plastome degeneration. In *L. kiusiana* and *L. japonica* plastomes, *psa* (5), *psb* (15), *pet* (6), *atp* (6), *rpo* (4) and *rbc*L were no longer evident, and the number of pseudogenes did not exceed three. Further loss of the chloroplast genome is followed by the deletion of other metabolic genes (such as *acc*D, *clp*P and *ycf*1/2), along with all other remaining housekeeping genes, including *trn*E. Thereafter, a stage of 'total deletion' is reached, in which the chloroplast genome is completely eradicated. At present, the data of mycoheterotrophic plastomes belonging to the Vanilloideae subfamily is limited, and it remains to be confirmed whether any species have undergone further plastome degeneration. However, it has been shown that the Vanilleae tribe has a very high evolutionary rate, and the plastome degeneration of mycoheterotrophic orchids in this tribe are various. Therefore, it is imperative to decode more chloroplast genome data of mycoheterotrophic plants belonging to the Vanilleae tribe.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: https://www.ncbi.nlm.nih.gov/nuccore/MW528436.1/.

Author contributions

LZ, SG and JL conceived the research and experimental design. JL collected the plant material. LZ, TC and XQ performed the experiments and analyzed the data. LZ, SG and JL designed the draft manuscript. All authors approved the final draft for submission and take full public responsibility for the content of the manuscript.

Funding

This research was supported by National Natural Science Foundation of China (No.81903749 and No. 81973426) and CAMS Innovation Fund for Medical Sciences (CIFMS) (No.2021-I2M-1-031).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1160446/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Codon frequencies and RSCU values of *G. lindleyana*: **(A)** Amino acid frequencies in protein-coding genes; **(B)** RSCU values of 20 amino acids and stop codons in 32 protein-coding genes.

References

Amiryousefi, A., Hyvönen, J., and Poczai, P. (2017). The plastid genome of vanillon (*Vanilla pompona*, orchidaceae). *Mitochondrial DNA Part B* 2 (2), 689–691. doi: 10.1080/23802359.2017.1383201

Amiryousefi, A., Hyvönen, J., and Poczai, P. (2018). IRscope: An online program to visualize the junction sites of chloroplast genomes. *Bioinformatics* 34, 3030–3031. doi: 10.1093/bioinformatics/bty220

Balakirev, E. S., and Ayala, F. J. (2003). Pseudogenes: are they "junk" or functional DNA? *Annu. Rev. Genet.* 37, 123–151. doi: 10.1146/annurev.genet.37.040103.103949

Barrett, C. F., and Davis, J. I. (2012). The plastid genome of the mycoheterotrophic *Corallorhiza striata* (Orchidaceae) is in the relatively early stages of degradation. *Am. J. Botany* 99, 1513–1523. doi: 10.3732/ajb.1200256

Barrett, C. F., Freudenstein, J. V., Li, J., Mayfield-Jones, D. R., Perez, L., Pires, J. C., et al. (2014a). Investigating the path of plastid genome degradation in early-transitional heterotrophic orchids, and implications for heterotrophic angiosperms. *Mol. Biol. Evol.* 31, 3095–3112. doi: 10.1093/molbev/msu252

Barrett, C. F., Specht, C. D., Jim, L. M., Wm, S. D., Zomlefer, W. B., and Davis, J. I. (2014b). Resolving ancient radiations: Can complete plastid gene sets elucidate deep relationships among the tropical gingers (Zingiberales)? *Ann. Botany* 113, 119–133. doi: 10.1093/aob/mct264

Beier, S., Thiel, T., Münch, T., Scholz, U., and Mascher, M. (2017). MISA-web: a web server for microsatellite prediction. *Bioinformatics* 33, 2583–2585. doi: 10.1093/bioinformatics/btx198

Bellot, S., and Renner, S. S. (2016). The plastomes of two species in the endoparasite genus *Pilostyles* (Apodanthaceae) each retain just fifive or six possibly functional genes. *Genome Biol. Evol.* 8, 189–201. doi: 10.1093/gbe/evv251

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi: 10.1093/nar/27.2.573

Bidartondo, M. I. (2005). The evolutionary ecology of myco-heterotrophy. *New Phytologist.* 167, 335–352. doi: 10.2307/3694504

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Bromham, L., Cowman, P. F., and Lanfear, R. (2013). Parasitic plants have increased rates of molecular evolution across all three genomes. *BMC Evol. Biol.* 13, 1–11. doi: 10.1186/1471-2148-13-126

Cameron, K. M. (2009). On the value of nuclear and mitochondrial gene sequences for reconstructing the phylogeny of vanilloid orchids (Vanilloideae, orchidaceae). *Ann. Bot.* 104 (3), 377–385. doi: 10.1093/aob/mcp024

Chang, C. C., Lin, H. C., Lin, I. P., Chow, T. Y., Chen, H. H., Chen, W. H., et al. (2006). 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. doi: 10.1093/molbev/msj029

Chase, M. W., Cameron, K. M., Freudenstein, J. V., Pridgeon, A. M., Salazar, G., Van den Berg, C., et al. (2015). An updated classification of orchidaceae. *Botanical J. Linn. Soc.* 177 (2), 151–174. doi: 10.1111/boj.12234

Christenhusz, M., and Byng, J. (2016). The number of known plant species in the world and its annual increase. *Phytotaxa* 261, 201–217. doi: 10.11646/phytotaxa.261.3.1

Chumley, T. W., Palmer, J. D., Mower, J. P., Fourcade, H. M., Calie, P. J., Boore, J. L., et al. (2006). 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.* 11, 2175–2190. doi: 10.1093/molbev/msl089

Delannoy, E., Fujii, S., Des Francs-Small, C. C., Brundrett, M., and Small, I. (2011). Rampant gene loss in the underground orchid *Rhizanthella gardneri* highlight evolutionary constraints on plastid genomes. *Mol. Biol. Evol.* 28, 2077–2086. doi: 10.1093/molbev/msr028

Downie, S. R., and Jansen, R. K. (2015). A comparative analysis of whole plastid genomes from the apiales: Expansion and contraction of the inverted repeat, mitochondrial to plastid transfer of DNA, and identification of highly divergent noncoding regions. *Syst. Bot.* 40, 336–351. doi: 10.1600/036364415X686620

Doyle, J. J., Doyle, J. L., Ballenger, J. A., and Palmer, J. D. (1996). The distribution and phylogenetic significance of a 50 kb chloroplast DNA inversion in the flowering plant family leguminosae. *Mol. Phylogenet. Evol.* 5, 429–438. doi: 10.1006/mpev.1996.0038

Feng, Y. L., Wicke, S., Li, J. W., Han, Y., Lin, C. S., Li, D. Z., et al. (2016). Lineage-specific reductions of plastid genomes in an orchid tribe with partially and fully mycoheterotrophic species. *Genome Biol. Evol.* 7, 2164–2175. doi: 10.1093/gbe/evw144

Frailey, D. C., Chaluvadi, S. R., Vaughn, J. N., Coatney, C. G., and Bennetzen, J. L. (2018). Gene loss and genome rearrangement in the plastids of five hemiparasites in the family orobanchaceae. *BMC Plant Biol.* 18, 30. doi: 10.1186/s12870-018-1249-x

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., and Dubchak, I. (2004). VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* 32 (suppl_2), W273–W279. doi: 10.1093/nar/gkh458

Funk, H. T., Berg, S., Krupinska, K., Maier, U. G., and Krause, K. (2007). Complete DNA sequences of the plastid genomes of two parasitic flowering plant species, *Cuscuta reflexa* and *Cuscuta gronovii*. *BMC Plant Biol.* 7, 45. doi: 10.1186/1471-2229-7-45

Gao, L., Yi, X., Yang, Y. X., Su, Y. J., and Wang, T. (2009). Complete chloroplast genome sequence of a tree fern *Alsophila spinulosa*: Insights into evolutionary changes in fern chloroplast genomes. *BMC Evol Biol.* 9, 130. doi: 10.1186/1471-2148-9-130

Goulding, S. E., Olmstead, R. G., Morden, C. W., and Wolfe, K. H. (1996). Ebb and flow of the chloroplast inverted repeat. *Mol. Gen. Genet.* 252, 195–206. doi: 10.1007/BF02173220

Graham, S. W., Lam, V. K. Y., and Merckx, V. S. F. T. (2017). Plastomes on the edge: The evolutionary breakdown of mycoheterotroph plastid genomes. *New Phytologist.* 214, 48–55. doi: 10.1111/nph.14398

Hirao, T., Watanabe, A., Kurita, M., Kondo, T., and Takata, K. (2008). Complete nucleotide sequence of the *Cryptomeria japonica* d. don. chloroplast genome and comparative chloroplast genomics: diversified genomic structure of coniferous species. *BMC Plant Biol.* 8, 70. doi: 10.1186/1471-2229-8-70

Jansen, R. K., and Palmer, J. D. (1987). Chloroplast DNA from lettuce and *Barnadesia* (Asteraceae): Structure, gene localization, and characterization of a large inversion. *Curr. Genet.* 11, 553–564. doi: 10.1007/BF00384619

Jansen, R. K., Wojciechowski, M. F., Sanniyasi, E., Lee, S. B., and Daniell, H. (2008). 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. doi: 10.1016/j.ympev.2008.06.013

Jin, J. J., Yu, W. B., Yang, J. B., Song, Y., Yi, T. S., and Li, D. Z. (2020). GetOrganelle: A fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. *Genome Biol.* 21, 241. doi: 10.1101/256479

Kang, M. J., Kim, S. C., Lee, H. R., Lee, S. A., Lee, J. W., Kim, T. D., et al. (2020). The complete chloroplast genome of Korean *Gastrodia elata* blume. *Mitochondrial DNA Part b-resources.* 19, 908–917. doi: 10.1080/23802359.2020.1721346

Katoh, K., Misawa, K., Kuma, K. I., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi: 10.1093/nar/gkf436

Kim, Y. K., Jo, S., Cheon, S. H., Joo, M. J., Hong, J. R., Kwak, M. H., et al. (2019). Extensive losses of photosynthesis genes in the plastome of a mycoheterotrophic orchid, *Cyrtosia septentrionalis* (Vanilloideae: Orchidaceae). *Genome Biol. Evol.* 11, 565–571. doi: 10.1093/gbe/evz024

Kim, Y. K., Jo, S. J., Cheon, S. H., Joo, M. J., Hong, J. R., Kwak, M., et al. (2020). Plastome evolution and phylogeny of orchidaceae, with 24 new sequences. *Front. Plant Sci.* 11, 1–11. doi: 10.3389/fpls.2020.00022

Kim, H. T., Kim, J. S., Moore, M. J., Neubig, K. M., Williams, N. H., Whitten, W. M., et al. (2015). Seven new complete plastome sequences reveal rampant independent loss of the *ndh* gene family across orchids and associated instability of the inverted repeat/small single-copy region boundaries. *PloS One* 10, e0142215. doi: 10.1371/journal.pone.0142215

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., et al. (2004). Versatile and open software for comparing large genomes. *Genome Biol.* 5, R12. doi: 10.1186/gb-2004-5-2-r12

Lam, V. K. Y., Gomez, M. S., and Graham, S. W. (2015). The highly reduced plastome of mycoheterotrophic *Sciaphila* (Triuridaceae) is colinear with its green relatives and is under strong purifying selection. *Genome Biol. Evol.* 7, 2220–2236. doi: 10.1093/gbe/evv134

Lam, V. K. Y., Merckx, V. S. F. T., and Graham, S. W. (2016). A few-gene plastid phylogenetic framework for mycoheterotrophic monocots. *Am. J. Bot.* 103, 692–708. doi: 10.3732/ajb

Leake, J. R. (1994). The biology of myco-heterotrophic ('saprophytic') plants. *New Phytol.* 127, 171–216. doi: 10.1111/j.1469-8137.1994.tb04272.x

Lee, H. L., Jansen, R. K., Chumley, T. W., and Kim, K. J. (2007). Gene relocations within chloroplast genomes of *Jasminum* and *Menodora* (Oleaceae) are due to multiple, overlapping inversions. *Mol. Biol. Evol.* 24, 1161–1180. doi: 10.1093/molbev/msm036

Li, M. H., Zhang, G. Q., Lan, S. R., and Liu, Z. J. (2016). A molecular phylogeny of Chinese orchids. *J. Syst. Evol.* 54, 349–362. doi: 10.1111/jse.12187

Lin, C. S., Chen, J. J. W., Huang, Y. T., Chan, M. T., Daniell, H., Chang, W. J., et al. (2015). The location and translocation of ndh genes of chloroplast origin in the orchidaceae family. *Sci. Rep.* 5, 1–10. doi: 10.1038/srep09040

Lin, C. P., Huang, J. P., Wu, C. S., Hsu, C. Y., and Chaw, S. M. (2010). Comparative chloroplast genomics reveals the evolution of pinaceae genera and subfamilies. *Genome Biol. Evol.* 2, 504–517. doi: 10.1093/gbe/evq036

Lohan, A. J., and Wolfe, K. H. (1998). A subset of conserved tRNA genes in plastid DNA of nongreen plants. *Genetics* 150, 425–433. doi: 10.1046/j.1365-2443.1998.00217.x

Luo, J., Hou, B. W., Niu, Z. T., Liu, W., Xue, Q. Y., and Ding, X. Y. (2014). Comparative chloroplast genomes of photosynthetic orchids: insights into evolution of the orchidaceae and development of molecular markers for phylogenetic applications. *PloS One* 9, e99016. doi: 10.1371/journal.pone.0099016

Merckx, V. (2013). *Mycoheterotrophy: the biology of plants living on fungi* (New York, NY, USA: Springer-Verlag). doi: 10.1007/978-1-4614-5209-6

Merckx, V., and Freudenstein, J. V. (2010). Evolution of mycoheterotrophy in plants: A phylogenetic perspective. *New Phytologist.* 185, 605–609. doi: 10.1111/j.1469-8137.2009.03155.x

Naumann, J., Der, J. P., Wafula, E. K., Jones, S. S., Wagner, S. T., Honaas, L. A., et al. (2016). Detecting and characterizing the highly divergent plastid genome of the nonphotosynthetic parasitic plant *Hydnora visseri* (Hydnoraceae). *Genome Biol. Evol.* 8, 345–363. doi: 10.1093/gbe/evv256

Niu, Z. T., Pan, J. J., Zhu, S. Y., Li, L. D., Xue, Q. Y., Liu, W., et al. (2017). Comparative analysis of the complete plastomes of *Apostasia wallichii* and *Neuwiedia singapureana* (Apostasioideae) reveals different evolutionary dynamics of IR/SSC boundary among photosynthetic orchids. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.01713

Park, J., Suh, Y., and Kim, S. (2020). A complete chloroplast genome sequence of *Gastrodia elata*(Orchidaceae) represents high sequence variation in the species. *Mitochondrial Part B-Resources* 5, 517–519. doi: 10.1080/23802359.2019.1710588

Petersen, G., Cuenca, A., Møller, M., and Seberg, O. (2015). Massive gene loss in mistletoe (*Viscum*, viscaceae) mitochondria. *Sci. Rep.* 5, 17588. doi: 10.1038/srep17588

Poliseno, L., Salmena, L., Zhang, J. W., Carver, B., Haveman, W. J., and Paolo, P. (2010). A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* 465, 1033–1038. doi: 10.1038/nature09144

Ravi, V., Khurana, J. P., Tyagi, A. K., and Khurana, P. (2008). An update on chloroplast genomes. *Plant Syst. Evol.* 271, 101–122. doi: 10.1007/s00606-007-0608-0

Roma, L., Cozzolino, S., Schlüter, P. M., Scopece, G., and Cafasso, D. (2018). The complete plastid genomes of *Ophrys iricolor* and *O. sphegodes* (Orchidaceae) and comparative analyses with other orchids. *PloS One* 13, e0204174. doi: 10.1371/journal.pone.0204174

Sanderson, M. J., Copetti, D., Búrquez, A., Busramante, E., Charboneau, J. L. M., Eguiarte, L. E., et al. (2015). Exceptional reduction of the plastid genome of saguaro cactus (*Carnegiea gigantea*): Loss of the *ndh* gene suite and inverted repeat. *Am. J. Botany* 102, 1115–1127. doi: 10.3732/ajb.1500184

Saski, C., Lee, S. B., Daniell, H., Wood, T. C., Tomkins, J., Kim, H. G., et al. (2005). Complete chloroplast genome sequence of *Glycine max* and comparative analyses with other legume genomes. *Plant Mol. Biol.* 59, 309–322. doi: 10.1007/s11103-005-8882-0

Sharp, P. M., and Li, W. H. (1987). The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15 (3), 1281–1295. doi: 10.1093/nar/15.3.1281

Shi, L. C., Chen, H. M., Jiang, M., Wang, L. Q., Wu, X., Huang, L. F., et al. (2019). CPGAVAS2, an integrated plastome sequence annotator and analyzer. *Nucleic Acids Res.* 47 (W1), W65–W73. doi: 10.1093/nar/gkz345

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Wang, H., Harrison, S. P., Prentice, I. C., Yang, Y., Bai, F., Togashi, H. F., et al. (2018). The China plant trait database: Toward a comprehensive regional compilation of functional traits for land plants. *Ecology* 99, 500. doi: 10.1002/ecy.2091

Weng, M. L., Blazier, J. C., Madhumita, G., and Jansen, R. K. (2014). Reconstruction of the ancestral plastid genome in geraniaceae reveals a correlation between genome rearrangements, repeats, and nucleotide substitution rates. *Mol. Biol. Evol.* 3, 645–659. doi: 10.1093/molbev/mst257

Westwood, J. H., Yoder, J. I., Timko, M. P., and Depamphilis, C. W. (2010). The evolution of parasitism in plants. *Trends Plant Sci.* 15, 227–235. doi: 10.1016/j.tplants.2010.01.004

Wicke, S., Müller, K. F., Depamphilis, C. W., Quandt, D., Wickett, N. J., Zhang, Y., et al. (2013). Mechanisms of functional and physical genome reduction in photosynthetic and nonphotosynthetic parasitic plants of the broomrape family. *Plant Cell.* 25, 3711–3725. doi: 10.1105/tpc.113.113373

Wicke, S., Müller, K. F., Depamphilis, C. W., and Schneeweiss, G. M. (2016). Mechanistic model of evolutionary rate variation en route to a nonphotosynthetic lifestyle in plants. *Proc. Natl. Acad. Sci.* 113, 9045–9050. doi: 10.1073/pnas.1607576113

Wicke, S., and Naumann, J. (2018). Molecular evolution of plastid genomes in parasitic flowering plants. *Adv. Botanical Res.* 85, 315–347. doi: 10.1016/bs.abr.2017.11.014

Wicke, S., Schaferhoff, B., Depamphilis, C. W., and Müller, K. F. (2014). Disproportional plastome-wide increase of substitution rates and relaxed purifying selection in genes of carnivorous lentibulariaceae. *Mol. Biol. Evol.* 31, 529–545. doi: 10.1093/molbev/mst261

Wicke, S., Schneeweiss, G. M., Depamphilis, C. W., Müller, K. F., and Quandt, D. (2011). The evolution of the plastid chromosome in land plants: Gene content, gene order, gene function. *Plant Mol. Biol.* 76, 273–297. doi: 10.1007/s11103-011-9762-4

Wolfe, K. H., Morden, C. W., and Palmer, J. D. (1992). Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. *Proc. Natl. Acad. Sci.* 89, 10648–10652. doi: 10.1073/pnas.89.22.10648

Wu, C. S., Lai, Y. T., Lin, C. P., Wang, Y. N., and Chaw, S. M. (2009). Evolution of reduced and compact chloroplast genomes (cpDNAs) in gnetophytes: Selection toward a lower-cost strategy. *Mol. Phylogenet. Evol.* 52, 115–124. doi: 10.1016/j.ympev.2008.12.026

Wu, C. S., Wang, Y. N., Hsu, C. Y., Lin, C. P., and Chaw, S. M. (2011). Loss of different inverted repeat copies from the chloroplast genomes of pinaceae and cupressophytes and influence of heterotachy on the evaluation of gymnosperm phylogeny. *Genome Biol. Evol.* 3, 1284–1295. doi: 10.1093/gbe/evr095

Yan, M. H., Moore, M. J., Meng, A. P., and Yao, X. H. (2017). The first complete plastome sequence of the basal asterid family styracaceae (Ericales) reveals a large inversion. *Plant Syst. Evol.* 303, 61–70. doi: 10.1007/s00606-016-1352-0

Yang, J. B., Tang, M., Li, H. T., Zhang, Z. R., and Li, D. Z. (2013). Complete chloroplast genome of the genus *Cymbidium*: Lights into the species identification, phylogenetic implications and population genetic analyses. *BMC Evol. Biol.* 13, 84. doi: 10.1186/1471-2148-13-84

Yang, M., Zhang, X. W., Liu, G. M., Yin, Y. X., Chen, K. F., Yun, Q. Z., et al. (2010). The complete chloroplast genome sequence of date palm (*Phoenix dactylifera* l.). *PloS One* 5, e12762. doi: 10.1371/journal.pone.0012762

Yuan, Y., Jin, X. H., Liu, J., Zhao, X., Zhou, J. H., Wang, X., et al. (2018). The *Gastrodia elata* genome provides insights into plant adaptation to heterotrophy. *Nat. Commun.* 9, 1615. doi: 10.1038/s41467-018-03423-5

Yue, F., Cui, L. Y., Depamphilis, C. W., Moret, B. M., and Tang, J. J. (2008). Gene rearrangement analysis and ancestral order inference from chloroplast genomes with inverted repeat. *BMC Genomics* 9, S25. doi: 10.1186/1471-2164-9-S1-S25

Zeng, S. Y., Zhou, T., Han, K., Yang, Y. C., Zhao, J. H., and Liu, Z. L. (2017). The complete chloroplast genome sequences of six *Rehmannia* species. *Genes* 8, 103. doi: 10.3390/genes8030103

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 09 May 2023
DOI 10.3389/fpls.2023.1179009

OPEN ACCESS

EDITED BY
Linchun Shi,
Chinese Academy of Medical Sciences and Peking Union Medical College, China

REVIEWED BY
Zhen-Hui Gong,
Northwest A&F University, China
Huie Li,
Guizhou University, China
Mingying Zhang,
Shaanxi University of Chinese Medicine, China

*CORRESPONDENCE
Xiaomin Wang
wangxiaomin_1981@163.com
Jianshe Li
13709587801@163.com

RECEIVED 03 March 2023
ACCEPTED 11 April 2023
PUBLISHED 09 May 2023

CITATION
Wang X, Bai S, Zhang Z, Zheng F, Song L, Wen L, Guo M, Cheng G, Yao W, Gao Y and Li J (2023) Comparative analysis of chloroplast genomes of 29 tomato germplasms: genome structures, phylogenetic relationships, and adaptive evolution.
Front. Plant Sci. 14:1179009.
doi: 10.3389/fpls.2023.1179009

Comparative analysis of chloroplast genomes of 29 tomato germplasms: genome structures, phylogenetic relationships, and adaptive evolution

Xiaomin Wang[1,2,3]*, Shengyi Bai[1], Zhaolei Zhang[4], Fushun Zheng[1], Lina Song[1], Lu Wen[1], Meng Guo[1,2,3], Guoxin Cheng[1,2,3], Wenkong Yao[1,2,3], Yanming Gao[1,2,3] and Jianshe Li[1,2,3]*

[1]College of Enology and Horticultrue, Ningxia University, Yinchuan, China, [2]Ningxia Modern Facility Horticulture Engineering Technology Research Center, Ningxia Facility Horticulture (Ningxia University) Technology Innovation Center, Yinchuan, China, [3]Key Laboratory of Modern Molecular Breeding for Dominant and Special Crops in Ningxia, Ningxia University, Yinchuan, China, [4]Hebei Key Laboratory of Study and Exploitation of Chinese Medicine, Chengde Medical University, Chengde, China

In order to compare and analyze the chloroplast (cp) genomes of tomato germplasms and understand their phylogenetic relationships, the cp genomes of 29 tomato germplasms were sequenced and analyzed in this study. The results showed highly conserved characteristics in structure, number of gene and intron, inverted repeat regions, and repeat sequences among the 29 cp genomes. Moreover, single-nucleotide polymorphism (SNP) loci with high polymorphism located at 17 fragments were selected as candidate SNP markers for future studies. In the phylogenetic tree, the cp genomes of tomatoes were clustered into two major clades, and the genetic relationship between *S. pimpinellifolium* and *S. lycopersicum* was very close. In addition, only *rps15* showed the highest average K_A/K_S ratio in the analysis of adaptive evolution, which was strongly positively selected. It may be very important for the study of adaptive evolution and breeding of tomato. In general, this study provides valuable information for further study of phylogenetic relationships, evolution, germplasm identification, and molecular marker-assisted selection breeding of tomato.

KEYWORDS

tomato, germplasm, chloroplast genome, phylogenetic, adaptive evolution

1 Introduction

Cultivated tomato (*Solanum lycopersicum*) is an annual or perennial herb, which is a model system for Solanaceae and fruiting vegetables. The origin center of tomato is the Andes Mountains of South America, which is native to Peru, Ecuador, and other places in South America (Aoki et al., 2010). It has become one of the major cultivated vegetables in the world and contains rich nutrients. Significant differences in the nutrient composition of different varieties were detected by using high-performance liquid chromatography and electrochemical methods (Wang et al., 2023). Previous studies have shown that *S. pimpinellifolium* is the ancestor of cultivated tomatoes (Lin et al., 2014b), and *S. habrochaites* is an important wild relative of cultivated tomato, which has a variety of excellent disease resistance and stress resistance traits (Safari et al., 2021). Wild germplasm resources are widely used in modern tomato breeding. Therefore, it is necessary to identify the phylogenetic relationships among tomato germplasms by chloroplast (cp) genome sequencing, assembly, and annotation.

Chloroplast is the main place for energy conversion and photosynthesis of plants (Sadali et al., 2019). Compared with the large nuclear genome, the cp genome is smaller, and the copy number is more (Ashworth and Vanessa, 2017). The genome of cp is generally maternal inheritance; there is no problem of gene recombination (Li et al., 2014; Zhao et al., 2015). In recent years, the cp genome has been mainly used in phylogenetic, population genetics, and phylogeography studies (Wang et al., 2022b), in which plant phylogenetic analysis is the most basic cp genome analysis, which has been widely used in the phylogenetic study of the plant kingdom or one of the groups and to identify the relationships of angiosperm order, family, genus, interspecies, and intraspecies (Sun et al., 2020; Guo et al., 2021; Li et al., 2021; Xie et al., 2021; Liu et al., 2022; Raman et al., 2022; Wang et al., 2022a). The cp genome has unique advantages in plant phylogenetic studies. It had been favored in the study of plant molecular systematics because of its distinctive differences of molecular evolution rates in different regions, moderate nucleic acid replacement rates, and easily accessible sequences. The cp genome contains a large number of functional genes related to photosynthesis, gene expression, and other biosynthesis. Most of the cytoplasmically inherited traits are maternally inherited, and the development of cp molecular markers associated with maternally inherited traits has important applications in molecular marker-assisted selection breeding. Studies have shown that yellowing is the most common mutant phenotype in Chinese cabbage (*Brassica campestris* ssp. *pekinensis*), which is mostly inherited maternally. The mutated gene is the cp 16S small subunit protein gene *rps4*, and the presence of a single-nucleotide polymorphism (SNP) (A–C) with a mutation rate higher than 99% in the coding region of *rps4* resulted in the conversion of the RPS4 protein's 193rd amino acid Val to Gly (Tang et al., 2018). Li et al. (2011) hypothesized that the upregulated expression of the RPS15a gene may be associated with the occurrence of the multi-ovary in wheat. Therefore, comparing cp genomes can identify some important variations in the evolution of species and provide a theoretical basis for the study of species relatedness and interspecific identification, while more cp molecular markers associated with maternal genetic traits can be developed for molecular marker-assisted selection breeding.

As an important family of plants, Solanaceae plants include many edible and medicinal plants (Martins da Silva et al., 2014), but there are only a few reports on the study of tomato cp genome. Daniell et al. (2006) conducted a comparative analysis of the complete cp genomes of wild potato (*Solanum bulbocastanum*), tomato (*S. lycopersicum*), tobacco (*Nicotiana tabacum*), and Atropa (*Atropa belladonna*). The results showed that deletions or insertions within some intergenic spacer regions result in less than 25% sequence identity. They can be used as an effective chloroplast marker in low-level phylogenetic studies. Chung et al. (2006) sequenced the cp genome of a cultivated potato (*Solanum tuberosum*) and compared it with the cp genome of six other Solanaceae plants, including *S. lycopersicum*. The results showed that there was a 241-bp deletion in the large single copy region (LSC) of cultivated potato, which could be used as a new method to identify cultivated potato and wild potato. Rachele et al. (2020) carried out a cp genome analysis with tomato as main material at first time. They sequenced seven cp genomes of cultivated accessions from Southern Italy and two wild species among the closest (*S. pimpinellifolium*) and most distantly related (*S. neorickii*) species to cultivated tomatoes. In total, 11 tomato cp genome sequences were retrieved in GenBank for comparative analysis with the abovementioned set. Finally, they found that *S. pimpinellifolium* was the nearest ancestor of all cultivated tomatoes. The local materials were closely related to other cultivated tomatoes. However, the SNP loci that could be used for future research were not screened, and adaptive evolution analysis was not carried out in the abovementioned study. Moreover, the research on the complete genome sequencing and analysis of tomato core germplasms from China and wild tomatoes' cp genome has not been reported widely.

In view of the lack of classification and phylogenetic relationship between tomato interspecies and intraspecies, 29 tomato germplasms with different genetic backgrounds at home and abroad were screened from the core germplasms by our research group, and six wild resources were collected from Tomato Genetic Resource Center. The second-generation high-throughput sequencing technology was used to sequence the complete cp genome of tomato germplasms as well as assemble and annotate. Then, bioinformatics analysis was performed; a high-definition map of cp genome and a phylogenetic tree were constructed to clarify the phylogenetic relationship of tomato germplasms. Moreover, SNP loci with high polymorphism

were selected as candidate SNP markers, and adaptive evolution analysis was conducted. This study will provide valuable information for further phylogenetic relationships, evolution, germplasm identification, and molecular marker-assisted selection breeding of tomato.

2 Materials and methods

2.1 Plant materials and genomic DNA isolation

In this study, 29 tomato core collections were selected and cultivated in the experimental farm of Ningxia University (Table 1). The fresh and healthy leaves of 29 tomato germplasms were collected, and then the leaf tissue samples were frozen fresh at -80°C until DNA extraction. The total genomic DNA was extracted using the New Plant Genome Extraction Kit (DP320) (Tiangen, Beijing, China) according to the manufacturer's instructions. The purity and integrity of the genomic DNA samples were identified using 1% agarose gel electrophoresis. The concentration of genomic DNA samples was measured using a NanoDrop 2000C spectrophotometer (Thermo Scientific; Waltham, MA, USA).

2.2 Tomato genome sequencing, assembly, and annotation

The high-throughput sequencing of 29 tomato germplasms was completed by Berry Genomics Co., Ltd. After the DNA samples were qualified, the genomic DNA was randomly cut into 350-bp fragments by enzyme digestion. After terminal repair and poly A addition, the sequencing adapters were connected at both ends of the fragments. Lastly, the libraries were analyzed for size distribution using agarose gels and were quantified using real-time PCR. The clustering of the index-coded samples was performed on a cBot Cluster Generation System using Novaseq 6000 S4 Reagent Kit (Illumina) according to the manufacturer's instructions. After cluster generation, the DNA libraries were sequenced on Illumina NovaSeq 6000 platform, and 150-bp paired-end reads were generated.

2.3 Analyses of repetitive sequences

Repetitive sequences in the cp genome play a critical role in genome evolution and rearrangements. The simple sequence repeats (SSR) motifs were analyzed in the cp genome of 29 tomato germplasms using MISA v2.1 (Raman et al., 2022). The minimum repeat thresholds of 10, six, five, five, five, and five are for mono-, di-, tri-, tetra-, penta-, and hexa-nucleotide SSRs, respectively. Tandem repeats were analyzed using the TRF (Benson, 1999) software with default parameters (Liu et al., 2022). In addition, oligonucleotide repeat analysis of four types of repeats in the cp genome was carried out. The forward, reverse, complement, and palindromic repeats were detected using REPuter online software (Kurtz et al., 2001) with a minimum repeat size of 30 bp and 90% sequence identity (Hamming distance of 3).

2.4 Boundary regions and chloroplast genome sequence comparison

The connecting regions of IR-LSC and IR-SSC in the cp genomes of 29 tomato germplasms were compared by using IRscope online software (https://irscope.shinyapps.io/irapp/) (Amiryousefi et al., 2018). The mVISTA online software (http://genome.lbl.gov/vista/mvista/submit.shtml) was used to compare the cp genomics of 29 tomato germplasms (Frazer et al., 2004). The comparative analysis was carried out by using the shuffle-LAGAN mode in mVISTA, and the sequence alignment was visualized in an mVISTA plot. The size of the sliding window is set to 100, and the default values for minimum and maximum Y are 50% to 100%. Mauve V2.4.0 was used to compare the cp genomes of 29 tomato germplasms to determine the collinearity of cp genome structure and identify possible rearrangements (Katoh et al., 2019). We also calculated the nucleotide diversity (Pi) of 80 protein-coding genes and intergenic spacer regions among the 29 tomato germplasms (Raman et al., 2022).

2.5 Analysis of phylogenetic relationship

The coding sequences, intergenomic sequences, and the complete cp genomes of 29 tomato germplasms were selected to construct a phylogenetic tree, and *Solanum bulbocastanum* DQ347958 was selected as the outgroup. To analyze the phylogenetic relationship of tomato, alignments were used to construct the phylogenetic trees using the maximum likelihood (ML) method implemented in RAxML v8.2.12 with 1,000 bootstrap replicates (Nguyen et al., 2015).

2.6 Analysis of substitution rate

In this study, the complete cp genomes of 29 tomato germplasms were compared. By extracting the same specific protein-encoded DNA sequence and translating it into a protein sequence, protein sequence alignment was performed using ParaAT3.0 software, and then the nucleic acid alignment result corresponding to a codon was translated back according to the protein alignment result. After the homologous sequence alignment, KaKs_Calculator 3.0 software was used to calculate the synonymous (K_S) and nonsynonymous (K_A) substitution rates and K_A/K_S ratios (Zhang, 2022).

3 Results

3.1 General features of the tomato chloroplast genome

The raw data were deposited in the Sequence Read Archive under BioProject accession number PRJNA936910. In this study, the cp genomes of 29 tomato germplasms were sequenced and

10.3389/fpls.2023.1179009

TABLE 1 Material sources of 29 tomato germplasms.

Sample name	Material name	Taxon	GenBank accession number	Location	Germplasm type	Mature fruit color	Fruit types
A1	LA3432	*S. lycopersicum*	OQ473528	USA	Cultivar	Red	Big fruit
A2	LA3474	*S. lycopersicum*	OQ473535	USA	Cultivar	Red	Big fruit
A3	LA2822	*S. lycopersicum*	OQ473542	USA	Cultivar	Red	Big fruit
A4	LA1969	*S. chilense*	OQ473544	Peru	Wild species	Red	Small fruit
A5	LA1269	*S. pimpinellifolium*	OQ473545	Peru	Wild species	Red	Small fruit
A6	20CL1036	*S. lycopersicum*	OQ473546	China	Cultivar	Red	Big fruit
A7	21CL0625	*S. lycopersicum*	OQ473547	China	Cultivar	Yellow	Cherry fruit
A8	21CL0668	*S. lycopersicum*	OQ473548	China	Cultivar	Pink	Big fruit
A9	21CL1999	*S. lycopersicum*	OQ473549	China	Cultivar	Red	Big fruit
A10	20CL2110	*S. lycopersicum*	OQ473521	China	Cultivar	Pink	Big fruit
A11	21CL0579	*S. lycopersicum*	OQ473522	China	Cultivar	Yellow	Cherry fruit
A12	21CL0031	*S. lycopersicum*	OQ473523	China	Cultivar	Pink	Cherry fruit
A14	21CL0381	*S. lycopersicum*	OQ473524	China	Cultivar	Red	Cherry fruit
A15	21CL0033	*S. lycopersicum*	OQ473525	China	Cultivar	Pink	Cherry fruit
A16	21CL1205	*S. lycopersicum*	OQ473526	China	Cultivar	Pink	Big fruit
A17	LA1245	*S. pimpinellifolium*	OQ473527	Ecuador	Wild species	Pink	Small fruit
A21	LA2399	*S. lycopersicum*	OQ473529	USA	Cultivar	Red	Big fruit
A23	D61825R	*S. lycopersicum*	OQ473530	China	Cultivar	Red	Big fruit
A24	D62333R	*S. lycopersicum*	OQ473531	China	Cultivar	Red	Big fruit
A27	D62140P	*S. lycopersicum*	OQ473532	China	Cultivar	Pink	Big fruit
A28	D62180P	*S. lycopersicum*	OQ473533	China	Cultivar	Pink	Big fruit
A29	D61867G	*S. lycopersicum*	OQ473534	China	Cultivar	Green	Big fruit
A33	Micro-tom	*S. lycopersicum*	OQ473536	USA	Cultivar	Red	Small fruit
A34	Moneymaker	*S. lycopersicum*	OQ473537	USA	Cultivar	Pink	Big fruit
A35	21CL2625	*S. lycopersicum*	OQ473538	China	Cultivar	Orange	Big fruit
A36	21CL2646	*S. lycopersicum*	OQ473539	China	Cultivar	Red	Big fruit
A38	LA2329	*S. habrochaites*	OQ473540	Peru	Wild species	Green	Small fruit
A39	LA2809	*S. peruvianum*	OQ473541	Peru	Wild species	Red	Small fruit
A41	62442	*S. habrochaites*	OQ473543	Peru	Wild species	Green	Small fruit

characterized. The final cp genomes were assembled, annotated, and submitted to GenBank. The complete cp genomes of the 29 tomato germplasms ranged from 155,257 to 155,461 bp in length. Each cp genome was made up of three distinct regions (Figure 1). The length of the IR region ranged from 25,594 to 25,612 bp. The length of the LSC region ranged from 85,688 bp (A39 and A41) to 85,875 bp (A4), and the length of the SSC region ranged from 18,355 bp (A38, A39, and A41) to 18,375 bp (A4). Among 29 tomato germplasms, the total GC content of the cp genomes of A38 was the lowest (37.84%), while that of A5 was the highest (37.87%). The total GC content of the cp genome of each of the remaining 27 germplasms was 37.86% (Supplementary Table S1).

In addition, the number of genes and introns were highly conserved, and the same suite of protein-coding genes, ribosomal

10.3389/fpls.2023.1179009

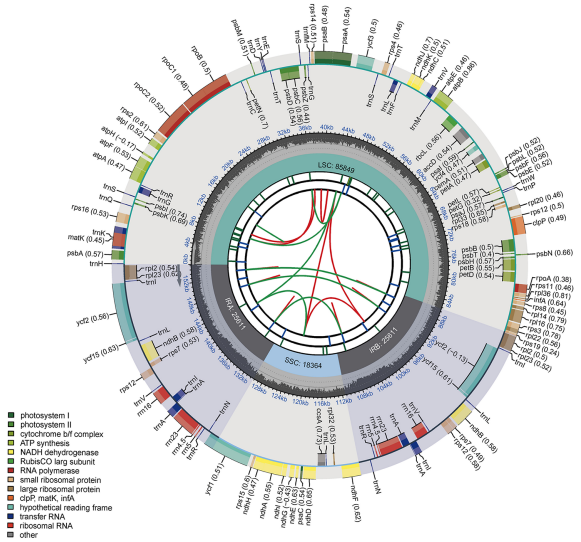


FIGURE 1
Circular chloroplast genome map of tomato. Genes drawn outside the circle are transcribed clockwise, and those inside are counterclockwise. Genes belonging to different functional groups are color-coded. The darker gray in the inner circle shows the GC content, while the lighter gray shows the AT content.

RNA (rRNA) genes, and transfer RNA (tRNA) genes was found in all taxa. Each cp genome included 113 unique genes, which contained 80 protein-coding genes, 29 tRNA genes, and four rRNA genes (Supplementary Table S2). In 113 genes, 18 genes with introns were identified. Among them, nine protein-coding (*atpF*, *ndhA*, *ndhB*, *petB*, *petD*, *rpl16*, *rpl2*, *rpoC1*, and *rps16*) and six tRNA genes (*trnA-UGC*, *trnG-GCC*, *trnI-GAU*, *trnK-UUU*, *trnL-UAA* and *trnV-UAC*) contained one intron, whereas *clpP*, *rps12*, and *ycf3* contained two introns.

3.2 Inverted repeat region contraction and expansion in chloroplast genomes

There were four borders between LSC, IRb, SSC, and IRa in the cp genome: LSC/IRb border, IRb/SSC border, SSC/IRa border, and IRa/LSC border. The borders of the 29 tomato germplasm cp genomes were compared (Figure 2). The four borders were conservative. The *rpl22* gene was present in the LSC region, and *rpl2* gene existed entirely in the IR region. Additionally, the *rps19* gene straddled the boundary of the LSC/IRb regions. The *ndhF* gene was located at the IRb/SSC border, and the distance between *ndhF* and the JSB line was 17 bp. The *ycf1* gene was observed at the JSA line, which straddled the boundary of the SSC/IRa regions. The *trnH* noncoding gene was located on the right side of the JLA line with a distance of 1 bp. In addition, the *psbA* gene existed in the LSC.

3.3 Codon usage in tomato chloroplast genomes

The amino acid frequency, codon usage, and relative synonymous codon usage (RSCU) of 80 protein-coding regions in 29 tomato germplasms were analyzed using Codon W. The RSCU values ranged from 63.99 to 64.03, the number of codons ranged from 23,010 to 23,016 in 29 tomato germplasms, and the number of amino acids ranged from 22,930 to 22,936. Of these amino acids, leucine (2,439–2,445 codons) was the most abundant amino acid, with a frequency of 10.63%–10.66%, while the frequency of cysteine (257 codons) was 1.12%. However, the most often used codon was ATT (encoding isoleucine), and the least used was TGA (termination codon). Almost all the amino acids had more than one synonymous codon; the exceptions were methionine and tryptophan. Furthermore, 62 codons displayed RSCU values exceeding 1.00. ATG and TGG, encoding methionine and tryptophan separately, exhibited no bias (RSCU = 1.00) (Supplementary Table S3). Moreover, three types of starting codons (ATG, GTG, and ACG) were detected in 80 protein-coding genes. Most genes used ATG as the starting codon. TAA, TAG, and TGA were present as stop codons in these genes. The most often used stop codon was TAA at 52.5%, followed by TAG (26.25%) and TGA (21.25%).

3.4 Comparative analysis of the repeat sequences in tomato genomes

In this study, the repeat sequences of the cp genomes of 29 tomato germplasms were analyzed. The distribution of SSRs, tandem repeats, and dispersed repeats differs among the 29 tomato germplasms (Figure 3A). Furthermore, we identified the total number of SSRs per cp genome that ranges from 36 to 42. Mononucleotides were the most frequent in the SSRs, with a distribution of 97.36%, followed by dinucleotide and trinucleotides at 2.47% and 0.17%, respectively, in the 29 tomato germplasms (Figure 3B). There was a dinucleotide with a

10.3389/fpls.2023.1179009

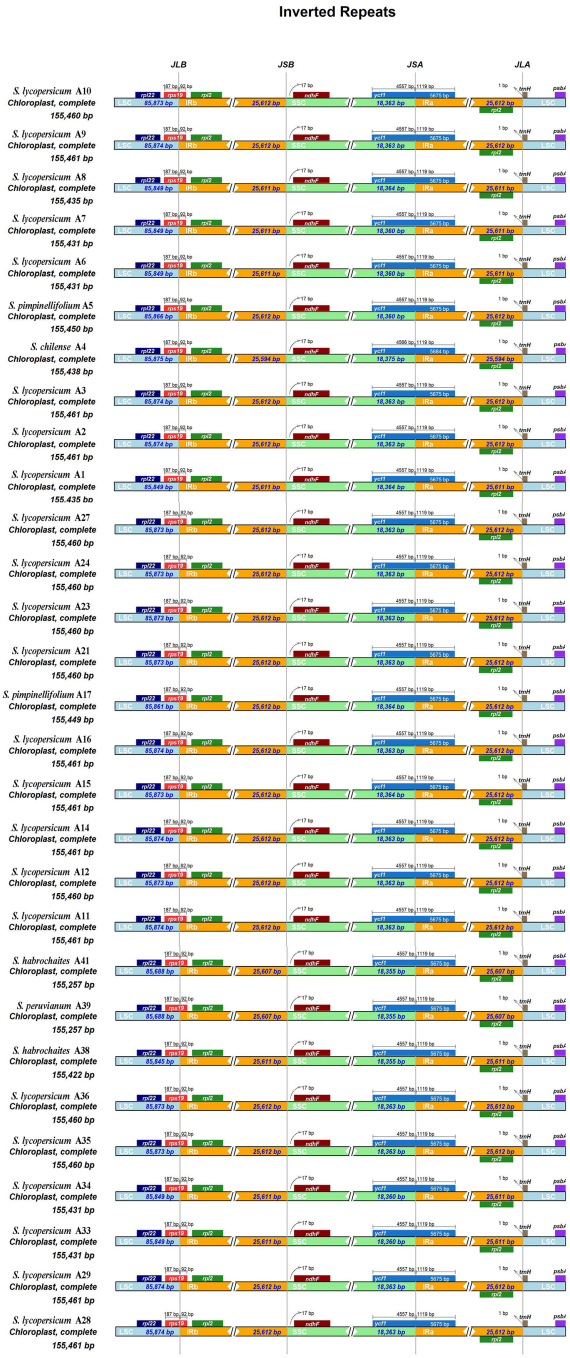


FIGURE 2
Comparison of the borders of large single copy (LSC), small single copy (SSC), and inverted repeat (IR) regions among 29 tomato germplasm chloroplast genomes. JLB line indicates the border between LSC and IRb, JSB line indicates the border between SSC and IRb, JSA line indicates the border between SSC and IRa, and JLA line indicates the border between LSC and IRa.

predominant motif of TA per cp genome in 29 tomato germplasms. Trinucleotides were absent in the cp genome of most tomato germplasms, except for a TTA motif in A4 and a TAA motif in A38. Moreover, the distribution of tandem repeats in the tomato cp genomes ranges from 26 to 30.

In addition to SSRs and tandem repeats, a dispersed repeat analysis of four types of repeats in the cp genome, including forward (F), reverse (R), palindromic (P), and complementary (C), was performed using REPuter. The total number of 39 dispersed repeats was identified in each cp genome of most tomato germplasms, while A4 and A5 had 40 dispersed repeats. Among the tomato cp genomes, forward repeats and palindromic repeats were the most common, accounting for 48.46% and 51.19%, respectively (Figure 3C). Only A4, A5, A39, and A41 had one reverse repeat, respectively. Complement repeats were not observed in the cp genomes of 29 tomato germplasms (Supplementary Table S4).

10.3389/fpls.2023.1179009

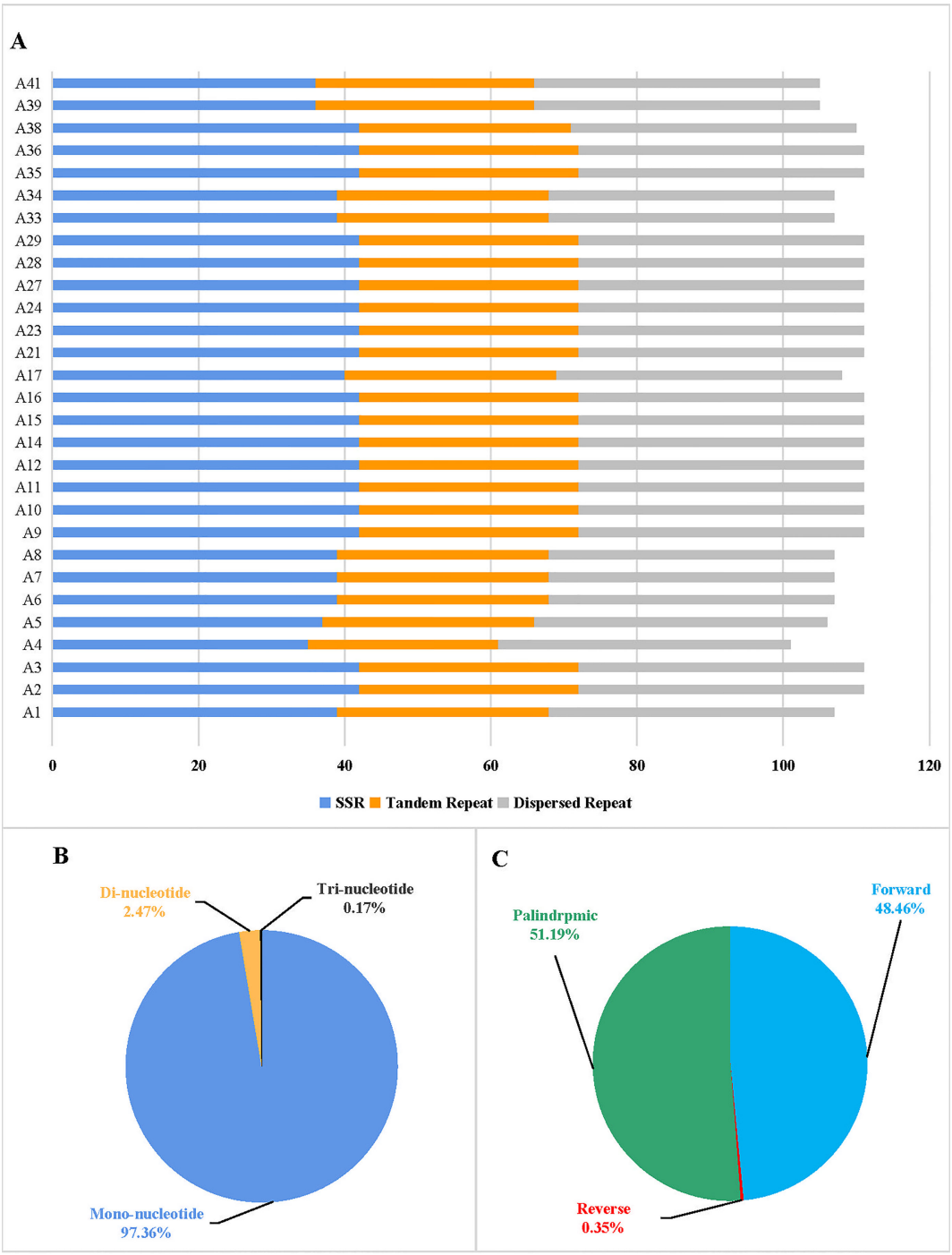


FIGURE 3
Histogram showing the number of repeats in 29 tomato germplasm chloroplast genomes. **(A)** Distribution of simple sequence repeats (SSRs), tandem repeats, and dispersed repeats in 29 tomato germplasms. **(B)** Proportion of different SSR repeat types. **(C)** Proportion of different dispersed repeat types.

3.5 Comparative chloroplast genome analysis

Multiple alignments of 29 tomato cp genomes were conducted using the online software platform mVISTA. A comparison of overall sequence variation showed that the tomato cp genome is highly conserved. Only *ycf1* open reading frame had a divergence in the coding region (Supplementary Figure S1). Subsequently, Mauve was used to identify the local collinear blocks (LCBs) of 29 tomato germplasm cp genomes (Supplementary Figure S2). These germplasms showed a consistent sequential order in all genes. The LCBs of all cp genomes showed a relatively high conservation with no gene rearrangement.

In addition, the nucleotide diversity (Pi) of 80 protein genes and intergenic spacer regions in the tomato cp genomes was calculated. A total of 80 genes were related to transcription, translation, and photosynthetic processes, with low variation (Figure 4A). Among these 80 genes, *petL* had the highest Pi value (0.0035), followed by *rps15* (0.0024) and *ycf1* (0.0015), showing obvious divergences. Meanwhile, 32 genes had a Pi value of 0. However, obvious divergences were detected in the intergenic spacer regions of *psbI-trnS-GCU*, *rpl36-infA*, *infA-rps8*, *trnR-ACG-trnN-GUU*, *ccsA-ndhD*, *ndhH-rps15*, *rps15-ycf1*, and *trnN-GUU-trnR-ACG* (Figure 4B). The Pi value of these six genes was above 0.003. In addition, SNP/MNP loci with high polymorphism located at 17 fragments were selected as candidate SNP markers for future studies (Table 2). Among the screened SNP markers, those localized to fragments of the *ndhH* gene and the *ndhK-ndhC-trnV-UAC* intergenic spacer regions could be used for interspecific identification.

3.6 Phylogenetic relationship analysis

In order to study the phylogenetic relationship among different tomato germplasms, ML phylogenetic trees were constructed using coding sequences, intergenomic sequences, and the complete cp genomes of 29 tomato germplasms (Figure 5). *Solanum bulbocastanum* DQ347958 selected as outgroups were retrieved from NCBI. In the phylogenetic tree, the tomatoes were clustered into two major clades based on coding sequences and the complete cp genomes, with one clade comprising all cultivated tomato and *S. pimpinellifolium* and the other clade containing the remaining four wild tomatoes. Furthermore, the first clade could be further divided into two minor clades, with all cultivated tomato clustered into the same minor clade and the other contained two *S. pimpinellifolium*, indicating a relatively closer interrelationship between cultivated tomato and *S. pimpinellifolium*. The second clade could also be divided into two minor clades, with one minor clade comprising only *S. chilense* and the other including *S. habrochaites* and *S. peruvianum* (Figures 5A, B). However, the phylogenetic results based on intergenomic sequences are different from the abovementioned results based on coding sequences and the complete cp genomes. Based on intergenomic sequences, the tomatoes were clustered into two major clades. The first can be further divided into two minor clades, with all cultivated tomatoes and two wild tomatoes (two *S. pimpinellifolium* and one *S. chilense*, respectively) clustered into the first minor clade and the other minor clade contained *S. habrochaites* A41 and *S. peruvianum* A39. *S. habrochaites* A38 was clustered separately as the second major clade. It indicated that *S. chilense* was more closely related to cultivated tomatoes than *S. habrochaites* and *S. peruvianum*.

3.7 Adaptive evolution analysis

In total, 49 protein-coding genes of all the 29 tomato cp genomes were used for the analysis of synonymous (K_S) and non-synonymous (K_A) substitution rates. The results showed that most protein-coding genes have relatively low average K_S values (<0.008), except the *petL* genes (Figure 6B). In the same way, the average K_A values of most protein-coding genes were comparatively low (<0.0015), except the *rps15* and *ycf1* genes (Figure 6A). The average K_A/K_S ratio of *rps15* gene was the highest (2.41). Furthermore, the K_A/K_S ratios of all the protein-coding genes ranged from 0 to 2.41, with an average ratio of only 0.14 (Figure 6C).

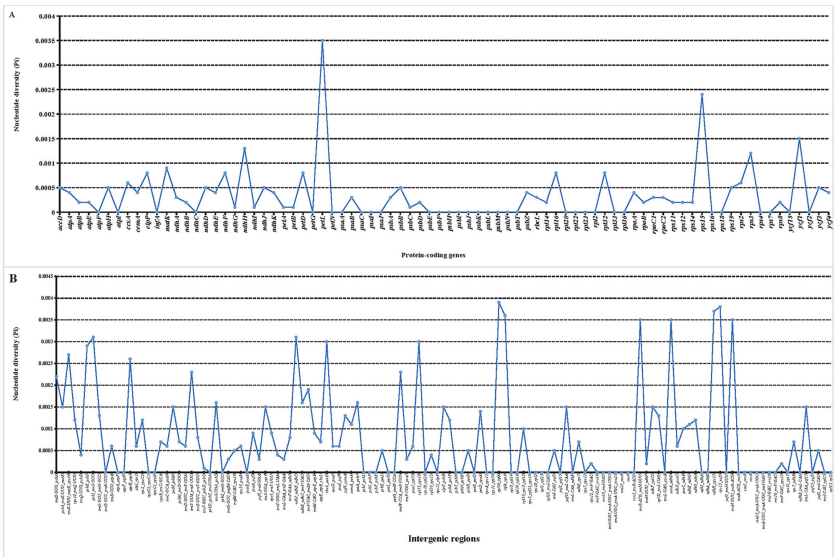


FIGURE 4
Nucleotide diversity (Pi) value in 29 tomato germplasm chloroplast genomes. **(A)** Pi value of protein-coding genes. **(B)** Pi value of intergenic spacer regions.

TABLE 2 Candidate polymorphic DNA markers among chloroplast genomes of 29 tomato germplasms.

No.	Position*	Polymorphic Type	Variant	Location
1	59,407-59,775	SNP	A/T, G/A, A/G, A/C, G/A	*accD*
2	111,655-113,611	SNP	G/T, A/G, C/T, G/A, G/A, C/A, C/G, G/A, C/T, C/T, A/G, C/T, A/C, C/A, C/A, G/A	*ndhF*
3	123,639-124,359	SNP	A/G, C/G, A/C, C/T, C/T, T/C, G/A	*ndhH*
4	84,531-85,109	SNP	A/G, G/A, A/G, C/T, G/A, C/A	*rps3*
5	17,372-18,226	SNP	C/T, G/A, T/C, G/A, A/C	*rpoC2*
6	24,194-24,416	SNP	G/T, C/T, C/A	*rpoB*
7	124,633-124,834	SNP	C/T, G/A, G/A, G/A	*rps15*
8	120,153-120,479	SNP	C/A, A/C, T/G	*ndhG-ndhI*
9	51,876-52,585	SNP	T/A, T/A, G/T, G/T, A/G, C/T, G/T, G/T	*ndhK-ndhC-trnV-UAC*
10	64,959-65,070	SNP	TA/AC, C/T, A/T, A/G, G/A, TG/AA, A/G, C/T, TT/CA, T/C, C/T, A/T, A/G, T/A, G/A, G/A	*petA-psbJ*
11	16,544-16,547	MNP	TTTC/AAAA	*rps2-rpoC2*
12	6330-7031	SNP/MNP	T/G, AT/GT, T/G, C/A, TTG/AAA	*rps16-trnQ-UUG*
13	58,197-58,228	SNP/MNP	A/T, T/C, C/T, TAGT/ACTA, A/G, G/A	*rbcL-accD*
14	49,422-49,995	SNP	C/A, C/A, A/G, G/A	*trnF-GAA-ndhJ*
15	208-437	SNP	G/A, A/G, T/G, C/A, G/A, G/A	*trnH-GUG-psbA*
16	109,378-109,612	MNP	CTTT/AAAG, AAA/TTT	*trnN-GUU-trnR-ACG*
17	131,673-131,907	MNP	TTT/AAA, AAAG/CTTT	*trnR-ACG-trnN-GUU*

*Position is based on the alignment file. SNP, single nucleotide polymorphism; MNP, multiple nucleotide polymorphism.

4 Discussion

Due to the slow evolution of plant cp genome, it has been used for plant classification and molecular evolution research. The study of species identification and phylogenetic evolution based on cp genome is a development trend of plant taxonomy biology, which has attracted more and more attention and recognition from researchers (Nock et al., 2015). The genetic background of tomato is more and more narrow due to the loss of genetic diversity, selfing characteristics, and long-term domestication in the process of tomato spreading from the origin to the world (Doebley et al., 2006). However, tomato wild relatives have a rich genetic diversity and are an inexhaustible gene library for the genetic improvement of tomato. So far, there are few reports on the comparative study of tomato cp genome. Rachele et al. (2020) compared and analyzed nine tomato cp genomes obtained by sequencing and 11 tomato cp genomes retrieved from GenBank. In this study, the cp genome sequences of 29 tomato germplasms were sequenced and compared, which also showed highly conserved characteristics in structure, number of gene and intron, inverted repeat regions, which was similar to the study of Rachele et al. (2020).

Repeats in the cp genome play an important role in genome evolution and rearrangements (Wang et al., 2022a). SSR can be used as a molecular marker and used in population genetics research because of its high polymorphism (Zhang et al., 2012; Zhao et al., 2015). However, among 29 tomato cp genomes, the number of mononucleotide repeats was the largest, accounting for the majority of all SSRs (97.36%). Therefore, polymorphisms existed in the SSRs of the tomato cp genome, but their repeat numbers were relatively conservative. Whether they could be used as molecular markers independently for population genetic analysis needs to be further validated. However, trinucleotide repeats were only present in A4 (*S. chilense*) and A38 (*S. habrochaites*) (TTA and TAA motifs, respectively), which may be used as SSR candidate markers for the study of subsequent germplasm identification. Moreover, the tandem repeats and dispersed repeats of tomato cp genome were relatively conservative. Previous studies have shown that repeats have a great influence on insertion and substitution, which can increase the genetic diversity of biological populations (Klein and O'Neill, 2018). The existence and abundance of cp repetitive sequences may also be related to multiple phylogenetic signals (Zhang et al., 2011; Wang et al., 2016), while the repetitive

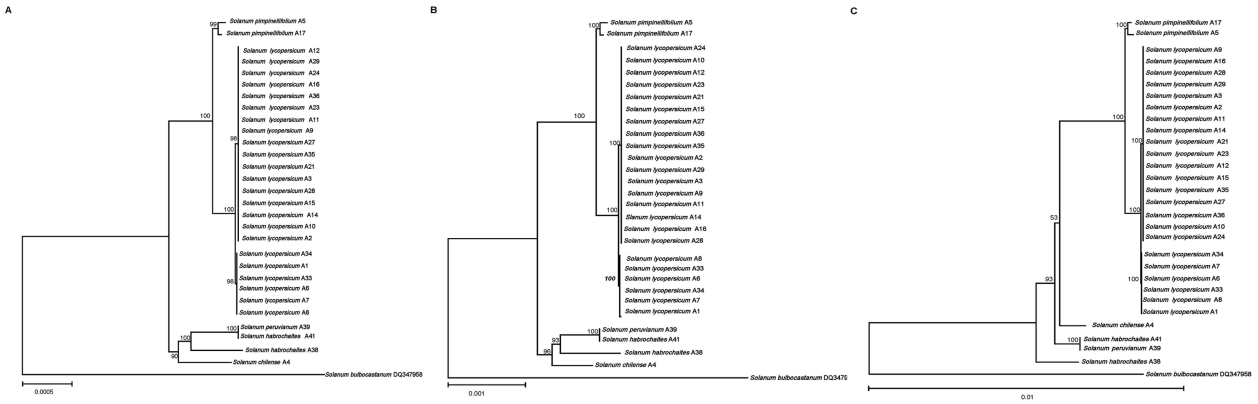


FIGURE 5
Maximum likelihood phylogenetic tree of 29 tomato germplasms. **(A)** Based on protein-coding sequences. **(B)** Based on complete chloroplast genomes. **(C)** Based on intergenomic sequences. Values above the branches represent the maximum likelihood bootstrap value. .

sequences on the tomato cp genome are relatively conservative, and whether it is related to phylogenetic signals needs further study.

In addition, 17 SNP/MNP loci with high polymorphism were selected as candidate SNP markers in this study. SNP molecular markers play an important role in tomato breeding, mainly in the identification of genetic relationship and genetic diversity of germplasm resources, the construction of genetic linkage map, the localization of target genes, and the identification of variety purity and molecular marker-assisted selection breeding (Kim et al., 2021). At present, the development of chloroplast SNP markers has been applied to many plants—for example, Cesare et al. (2010) identified six cp markers containing both cp SSRs and SNPs, and these SNP markers can distinguish most *Miscanthus* species and detect intraspecific variations, which can be used for breeding purposes. Lin et al. (2014a) developed four cp SNP molecular markers to distinguish soybean male sterile lines and maintainer lines and also to distinguish hybrids and soybean maintainer lines. The highly polymorphic SNP loci screened in this study were located at the regions of *ndhG-ndhI*, *ndhK-ndhC*, *petA-psbJ*, *rps2-rpoC2*, *rps16-trnQ-UUG*, *rbcL-accD*, *trnF-GAA-ndhJ*, *trnH-GUG-psbA*, *trnN-GUU-trnR-ACG*, and *trnR-ACG-trnN-GUU* and in *accD*, *ndhF*, *ndhH*, *rps3*, *rpoC2*, *rpoB*, and *rps15* genes. They will be developed as candidate cp genome SNP molecular markers for future studies. Among the screened SNP markers, those localized to segments of the *ndhH* gene and the *ndhK-ndhC-trnV-UAC* gene spacer region could be used for interspecific identification. The other developed SNP marker can be used to analyze genetic diversity and population structure at the cp genome level and to develop functional markers

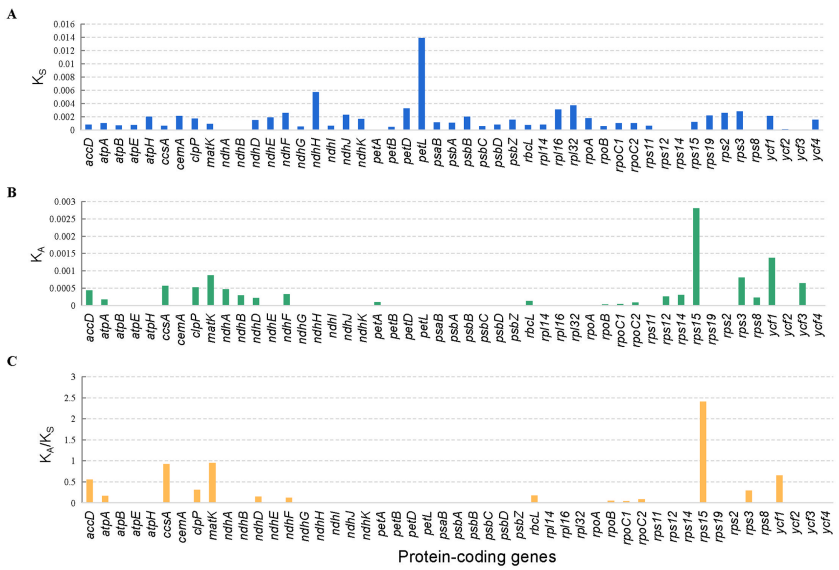


FIGURE 6
Selective pressure of 49 protein-coding genes in 29 tomato cp genomes. **(A)** K_S, rate of synonymous substitution. **(B)** K_A, rate of nonsynonymous substitution. **(C)** K_A/K_S, rate of non-synonymous *vs*. synonymous substitution.

associated with traits such as male sterility, which has an important application value in tomato germplasm identification and molecular marker-assisted selection breeding.

The cultivated tomatoes were domesticated from wild tomatoes. The fruit weight of modern cultivated tomato is more than 100 times that of its ancestors. In order to reveal the secret of tomatoes from small to large, Lin et al. (2014b) used genome-wide variation group data to analyze the phylogeny and population structure of tomatoes, and they found that the tomato population was divided into three subgroups, namely, *S. pimpinellifolium*, *S. lycopersicum* var. *cerasiforme*, and large-fruit cultivated tomato (*S. lycopersicum*). Combined with abundant phenotypic data and population genetics analysis, it was proved that wild tomato (*S. pimpinellifolium*) evolved into cherry tomato (*S. lycopersicum* var. *cerasiforme*) and finally formed a two-step artificial selection process of large-fruit cultivated tomato—namely, domestication and improvement. The comparative analysis of the 29 tomato germplasm cp genomes sequenced in this work allowed the phylogenetic relationships among wild and cultivated germplasms to be defined and also indicated that the genetic relationship between *S. pimpinellifolium* and *S. lycopersicum* was very close—that is, *S. pimpinellifolium* may be the ancestor of cultivated tomatoes—which was consistent with the previous research results. Notably, the phylogenetic relationships based on coding sequences and the complete cp genomes were consistent with the results based on traditional botanical classification (Zhao, 2012). However, the phylogenetic relationships based on intergenomic sequences were different from the results based on coding sequences and the complete cp genomes and traditional botanical classification, which may be due to the fact that intergenomic sequences contain more variability. The phylogenetic relationships based on intergenomic sequences show that *S. chilense* is more closely related to cultivated tomatoes and *S. pimpinellifolium* than *S. habrochaites* and *S. peruvianum*.

The K_A/K_S ratio is associated with gene adaptive evolution, such as positive selection and purity selection (Raman et al., 2022). In general, non-synonymous substitutions can cause amino acid changes, resulting in changes in protein conformation and function. Therefore, it will cause adaptive evolution and bring the advantages or disadvantages of natural selection. Synonymous substitution does not change the composition of the protein, so it is not affected by natural selection; then, Ks can reflect the background base substitution rate of the evolutionary process (Hurst, 2002). The K_A/K_S ratio can explain the type of selection of this gene. When K_A/K_S <<1, the gene is selected by purification. The K_A/K_S of most genes is far less than 1 because generally non-synonymous substitutions bring evolutionary disadvantages, and only a few cases will result in evolutionary advantages. When K_A/K_S >>1, the genes are strongly positively selected, and these genes are rapidly evolving recently and are of great significance for the evolution of species. We can screen some genes for further functional studies according to the K_A/K_S ratio, which has been widely applied to the field of molecular evolution (Navarro and Barton, 2003). In this study, the analysis results of synonymous (K_S) and non-synonymous (K_A) substitution rates showed that *petL* genes have relatively high average K_S values. The *rps15* and *ycf1* genes have relatively high average K_A. Only *rps15* showed the highest average K_A/K_S ratio of 2.41, which was strongly positively selected. It may be a gene that is rapidly evolving recently, and its function can be further studied. It may be very important for the study of adaptive evolution and breeding of tomato.

5 Conclusion

In this study, the latest sequencing results of the chloroplast genomes of 29 tomato germplasms were reported and compared. Genome annotation and comparative analysis showed that each chloroplast genome was a typical tetragonal structure. The 29 chloroplast genomes are highly conserved in terms of structure, gene and intron number, IR region, and repeat sequences. In addition, we had screened SNP/MNP loci with high polymorphism located at 17 fragments of the regions of *ndhG-ndhI*, *ndhK-ndhC*, *petA-psbJ*, *rps2-rpoC2*, *rps16-trnQ-UUG*, *rbcL-accD*, *trnF-GAA-ndhJ*, *trnH-GUG-psbA*, *trnN-GUU-trnR-ACG*, and *trnR-ACG-trnN-GUU* and in *accD*, *ndhF*, *ndhH*, *rps3*, *rpoC2*, *rpoB*, and *rps15* genes in the cp genomes of 29 tomato germplasms, which will be used as candidate SNP markers in future studies. In the phylogenetic tree, the cp genomes of tomatoes were clustered into two major clades, and the genetic relationship between *S. pimpinellifolium* and *S. lycopersicum* was very close. Moreover, in the analysis of adaptive evolution, only *rps15* showed the highest average K_A/K_S ratio of 2.4, which was strongly positively selected. In general, this study will provide valuable information for further study of phylogenetic relationships, germplasm identification, and molecular marker-assisted selection breeding of tomato.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Author contributions

XW and JL conceived and designed the study. FZ, LS, and LW collected and identified the plant materials. SB and ZZ performed the experiments and analyzed the data. XW and SB wrote the manuscript. XW, WY, MG, GC, and YG revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by Ningxia Hui Autonomous Region Agricultural Special and Dominant Industry Breeding Project (NXNYYZ20200104).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1179009/full#supplementary-material

References

Amiryousefi, A., Hyvönen, J., and Poczai, P. (2018). IRscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics* 34, 3030–3031. doi: 10.1093/bioinformatics/bty220

Aoki, K., Yano, K., Suzuki, A., Kawamura, S., Sakurai, N., Suda, K., et al. (2010). Large-Scale analysis of full-length cDNAs from the tomato *(Solanum lycopersicum)* cultivar micro-tom, a reference system for the solanaceae genomics. *BMC Genomics* 11, 210. doi: 10.1186/1471-2164-11-210

Ashworth,, and Vanessa, E. T. M. (2017). Revisiting phylogenetic relationships in phoradendreae (viscaceae): utility of the trnl-f region of chloroplast dna and presence of a homoplasious inversion in the intergenic spacer. *Botany* 95 (3), 1–12. doi: 10.1139/cjb-2016-0241

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi: 10.1093/nar/27.2.573

Cesare, M. D., Hodkinson, T. R., and Barth, S. (2010). Chloroplast DNA markers (cpSSRs, SNPs) for *Miscanthus*, *Saccharum* and related grasses (panicoideae, poaceae). *Mol. Breeding* 26 (3), 539–544. doi: 10.1007/s11032-010-9451-z

Chung, H. J., Jung, J. D., Park, H. W., Kim, J. H., Cha, H. W., Min, S. R., et al. (2006). The complete chloroplast genome sequences of solanum tuberosum and comparative analysis with solanaceae species identified the presence of a 241bp deletion in cultivated potato chloroplast DNA sequence. *Plant Cell Rep.* 25 (12), 1369–1379. doi: 10.1007/s00299-006-0196-4

Daniell, H., Lee, S. B., Grevich, J., Saski, C., Quesada-Vargas, T., and Guda, C. (2006). Complete chloroplast genome sequences of solanum bulbocastanum, solanum lycopersicum and comparative analyses with other solanaceae genomes. *Theor. Appl. Genet.* 112 (8), 1503–1518. doi: 10.1007/s00122-006-0254-x

Doebley, J. F., Gaut, B. S., and Smith, B. D. (2006). The molecular genetics of crop domestication. *Cell.* 127, 1309–1321. doi: 10.1016/j.cell.2006.12.006

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., and Inna, D. (2004). VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* 32, W273–W279. doi: 10.1093/nar/gkh458

Guo, X., Wang, Z., Zhang, Y., and Wang, R. (2021). Chromosomal-level assembly of the *Leptodermis oblonga* (rubiaceae) genome and its phylogenetic implications. *Genomics* 113 (5), 3072–3082. doi: 10.1016/j.ygeno.2021.07.012

Hurst, L. D. (2002). The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.* 18 (9), 486–487. doi: 10.1016/S0168-9525(02)02722-1

Katoh, K., Rozewicki, J., and Yamada, K. D. (2019). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform.* 20, 1160–1166. doi: 10.1093/bib/bbx108

Kim, M., Jung, J. K., Shim, E. J., Chung, S. M., Park, Y., Lee, G. P., et al. (2021). Genome-wide SNP discovery and core marker sets for DNA barcoding and variety identification in commercial tomato cultivars. *Scientia Horticulturae* 276, 109734. doi: 10.1016/j.scienta.2020.109734

Klein, S. J., and O'Neill, R. J. (2018). Transposable elements: genome innovation, chromosome diversity, and centromere conflict. *Chromosome Res.* 26 (1), 5–23. doi: 10.1007/s10577-017-9569-5

Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* 29, 4633–4642. doi: 10.1093/nar/29.22.4633

Li, X., Ma, S., Zhang, G., and Niu, N. (2011). Cloning of ribosomal protein S15a gene *(TaRPS15a)* and its expression patterns based on temporal-spatial in multi-ovary line of wheat *(Triticum aestivum)*. *J. Agric. Biotechnol.* 19 (2), 236–242. doi: 10.3969/j.issn.1674-7968.2011.02.006

Li, X., Yang, Y., Henry, R. J., Rossetto, M., Wang, Y., and Chen, S. (2014). Plant DNA barcoding: from gene to genome. *Biol. Rev.* 90 (1), 157–166. doi: 10.1111/brv.12104

Li, Q., Yu, Y., Zhang, Z., and Wen, J. (2021). Comparison among the chloroplast genomes of five species of *Chamaerhodos* (rosaceae: potentilleae): phylogenetic implications. *Nordic J. Botany* 39 (6), e03121. doi: 10.1111/njb.03121

Lin, C., Zhang, C., Zhao, H., Xing, S., Wang, Y., Liu, X., et al. (2014a). Sequencing of the chloroplast genomes of cytoplasmic male-sterile and male-fertile lines of soybean and identification of polymorphic markers. *Plant Science* 229, 208–214. doi: 10.1016/j.plantsci.2014.09.005

Lin, T., Zhu, G., Zhang, J., Xu, X., Yu, Q., Zheng, Z., et al. (2014b). Genomic analyses provide insights into the history of tomato breeding. *Nat. Genet.* 46, 1220–1226. doi: 10.1038/ng.3117

Liu, J., Shi, M., Zhang, Z., Xie, H., Kong, W., Wang, Q., et al. (2022). Phylogenomic analyses based on the plastid genome and concatenated nrDNA sequence data reveal cytonuclear discordance in genus *Atractylodes* (Asteraceae: carduoideae). *Front. Plant Science* 13. doi: 10.3389/fpls.2022.1045423

Martins da Silva, B. J., Rodrigues, A. P. D., Farias, L. H. S., Hage, A. A. P., Nascimento, J. L. M., and Silva, E. O. (2014). Physalis angulata induces *in vitro* differentiation of murine bone marrow cells into macrophages. *BMC Cell Biol.* 15. doi: 10.1186/1471-2121-15-37

Navarro, A., and Barton, N. H. (2003). Chromosomal speciation and molecular divergence-accelerated evolution in rearranged chromosomes. *Science* 300 (5617), 321–324. doi: 10.1126/science.1088277

Nguyen, L., Schmidt, H. A., Haeseler, A. V., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Nock, C. J., Waters, D., Edwards, M. A., Bowen, S. G., and Henry, R. J. (2015). Chloroplast genome sequences from total DNA for plant identification. *Plant Biotechnol. J.* 9 (3), 328–333. doi: 10.1111/j.1467-7652.2010.00558.x

Rachele, T., Lorenza, S., Donata, C., Concita, C., Luigi, O., Teodoro, C., et al. (2020). Cultivated tomato *(Solanum lycopersicum* l.) suffered a severe cytoplasmic bottleneck during domestication: implications from chloroplast genomes. *Plants* 9 (11), 1443. doi: 10.3390/plants9111443

Raman, G., Nam, G.-H., and Park, S. (2022). Extensive reorganization of the chloroplast genome of *Corydalis platycarpa*: a comparative analysis of their organization and evolution with other *Corydalis* plastomes. *Front. Plant Science.* 13. doi: 10.3389/fpls.2022.1043740

Sadali, N. M., Sowden, R. G., Ling, Q., and Jarvis, R. P. (2019). Differentiation of chromoplasts and other plastids in plants. *Plant Cell Rep.* 38 (7), 803–818. doi: 10.1007/s00299-019-02420-2

Safari, Z. S., Ding, P., Nakasha, J. J., and Yusoff, S. F. (2021). Controlling fusarium oxysporum tomato fruit rot under tropical condition using both chitosan and vanillin. *Coatings* 11, 367. doi: 10.3390/coatings11030367

Sun, C., Chen, F., Teng, N., Xu, Y., and Dai, Z. (2020). Comparative analysis of the complete chloroplast genome of seven *Nymphaea* species. *Aquat. Botany* 170, 103353. doi: 10.21203/rs.3.rs-20050/v1

Tang, X., Wang, Y., Zhang, Y., Huang, S., Liu, Z., Fei, D., et al. (2018). A missense mutation of plastid *RPS4* is associated with chlorophyll deficiency in Chinese cabbage *(Brassica campestris ssp. pekinensis)*. *BMC Plant Biol.* 18 (1), 130. doi: 10.1186/s12870-018-1353-y

Wang, W., Chen, S., and Zhang, X. (2016). Chloroplast genome evolution in actinidiaceae: *clpP* loss, heterogenous divergence and phylogenomic practice. *PloS One* 11 (9), e0162324. doi: 10.1371/journal.pone.0162324

Wang, J., Qian, J., Jiang, Y., Chen, X., Zheng, B., Chen, S., et al. (2022a). Comparative analysis of chloroplast genome and new insights into phylogenetic relationships of *Polygonatum* and tribe polygonateae. *Front. Plant Science* 13. doi: 10.3389/fpls.2022.882189

Wang, Y., Wen, F., Hong, X., Li, Z., Mi, Y., and Zhao, B. (2022b). Comparative chloroplast genome analyses of *Paraboea* (Gesneriaceae): insights into adaptive evolution and phylogenetic analysis. *Front. Plant Science* 13. doi: 10.3389/fpls.2022.1019831

Wang, J., Xu, Q., Liu, J., Kong, W., and Shi, L. (2023). Electrostatic self-assembly of MXene on ruthenium dioxide-modified carbon cloth for electrochemical detection of kaempferol. *Small* doi: 10.1002/smll.202301709

Xie, X., Huang, R., Li, F., Tian, E., Li, C., and Chao, Z. (2021). Phylogenetic position of *Bupleurum sikangense* inferred from the complete chloroplast genome sequence. *Gene.* 798, 145801. doi: 10.1016/j.gene.2021.145801

Zhang, Z. (2022). Kaks_calculator 3.0: calculating selective pressure on coding and non-coding sequences. *Genomics Proteomics Bioinf.* 20 (3), 536–540. doi: 10.1016/j.gpb.2021.12.002

Zhang, Q., Li, J., Zhao, Y., Korban, S., and Han, Y. (2012). Evaluation of genetic diversity in Chinese wild apple species along with apple cultivars using SSR markers. *Plant Mol. Biol. Rep.* 30 (3), 539–546. doi: 10.1007/s11105-011-0366-6

Zhang, Y., Ma, P., and Li, D. (2011). High-throughput sequencing of six bamboo chloroplast genomes: phylogenetic implications for temperate woody bamboos (Poaceae: bambusoideae). *PloS One* 6 (5), e20596. doi: 10.1371/journal.pone.0020596

Zhao, L. (2012). *Tomato wild resources* (Shanghai, China: Shanghai Jiaotong University Press).

Zhao, Y., Yin, J., Guo, H., Zhang, Y., Xiao, W., Sun, C., et al. (2015). The complete chloroplast genome provides insight into the evolution and polymorphism of panax ginseng. *Front. Plant Science* 5. doi: 10.3389/fpls.2014.00696

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 22 June 2023
DOI 10.3389/fpls.2023.1163325

OPEN ACCESS

EDITED BY
Weijun Kong,
Capital Medical University, China

REVIEWED BY
Wei Gu,
Nanjing University of Chinese Medicine, China
Rui He,
Guangzhou University of Chinese Medicine, China
Xilong Zheng,
Chinese Academy of Medical Sciences and Peking Union Medical College, China

*CORRESPONDENCE
Jingying Chen
cjy6601@163.com

RECEIVED 10 February 2023
ACCEPTED 17 May 2023
PUBLISHED 22 June 2023

CITATION
Zhang W, Zhang Z, Liu B, Chen J, Zhao Y and Huang Y (2023) Comparative analysis of 17 complete chloroplast genomes reveals intraspecific variation and relationships among *Pseudostellaria heterophylla* (Miq.) Pax populations. *Front. Plant Sci.* 14:1163325. doi: 10.3389/fpls.2023.1163325

Comparative analysis of 17 complete chloroplast genomes reveals intraspecific variation and relationships among *Pseudostellaria heterophylla* (Miq.) Pax populations

Wujun Zhang[1], Zhaolei Zhang[2], Baocai Liu[1], Jingying Chen[1]*, Yunqing Zhao[1] and Yingzhen Huang[1]

[1]Institute of Agricultural Bioresources, Fujian Academy of Agricultural Sciences, Fuzhou, China, [2]Hebei Key Laboratory of Study and Exploitation of Chinese Medicine, Chengde Medical University, Chengde, China

Pseudostellaria heterophylla (Miq.) Pax is a well-known medicinal and ecologically important plant. Effectively distinguishing its different genetic resources is essential for its breeding. Plant chloroplast genomes can provide much more information than traditional molecular markers and provide higher-resolution genetic analyses to distinguish closely related planting materials. Here, seventeen *P. heterophylla* samples from Anhui, Fujian, Guizhou, Hebei, Hunan, Jiangsu, and Shandong provinces were collected, and a genome skimming strategy was employed to obtain their chloroplast genomes. The *P. heterophylla* chloroplast genomes ranged from 149,356 bp to 149,592 bp in length, and a total of 111 unique genes were annotated, including 77 protein-coding genes, 30 tRNA genes, and four rRNA genes. Codon usage analysis showed that leucine had the highest frequency, while UUU (encoding phenylalanine) and UGC (encoding cysteine) were identified as the most and least frequently used codons, respectively. A total of 75–84 SSRs, 16–21 short tandem repeats, and 27–32 long repeat structures were identified in these chloroplast genomes. Then, four primer pairs were revealed for identifying SSR polymorphisms. Palindromes are the dominant type, accounting for an average of 47.86% of all long repeat sequences. Gene orders were highly collinear, and IR regions were highly conserved. Genome alignment indicated that there were four intergenic regions (*psaI-ycf4*, *ycf3-trnS*, *ndhC-trnV*, and *ndhI-ndhG*) and three coding genes (*ndhJ*, *ycf1*, and *rpl20*) that were highly variable among different *P. heterophylla* samples. Moreover, 10 SNP/MNP sites with high polymorphism were selected for further study. Phylogenetic analysis showed that populations of Chinese were clustered into a monophyletic group, in which the non-flowering variety formed a separate subclade with high statistical support. In this study, the comparative analysis of complete chloroplast

genomes revealed intraspecific variations in *P. heterophylla* and further supported the idea that chloroplast genomes could elucidate relatedness among closely related cultivation materials.

KEYWORDS

***Pseudostellaria heterophylla*, chloroplast genome, comparative analysis, intraspecific variation, phylogenetic relationship**

Introduction

Pseudostellaria heterophylla (Miq.) Pax (tai-zi-shen or hai-er-shen) is a well-known traditional medicinal plant of the Caryophyllaceae family. It is commonly used for the treatment of fatigue, spleen asthenia, anorexia, asthenia after severe illness, and cough due to lung dryness either in China (Commission, 2015) or in Korea (Liu et al., 2017). Recent pharmacologic research has indicated that *P. heterophylla* has anti-diabetes (Liu et al., 2017), immune enhancement (Yang et al., 2020), and anti-oxidant properties (Ng et al., 2004) due to its composition containing numerous active compounds such as cyclic peptides (pseudostellarin), polysaccharides, amino acids, saponins, and sapogenins (Wang et al., 2013). *P. heterophylla* is mainly distributed in the Fujian, Guizhou, Shandong, Anhui, and Jiangsu provinces of China (Kang et al., 2016), Japan, Korea, and the Russian Far East (Choi and Pak, 2000). *P. heterophylla* has been cultivated in China for over 100 years with abundant germplasm resources (Xiao et al., 2015), represented by significant variability in leaf length, leaf width, number of main stems, total biomass, and number of above-ground stem nodes. Currently, the breeding of *P. heterophylla* is progressing slowly since the introduction of varieties is not standardized and the genetic background of the cultivated populations cannot be traced. Moreover, there are few sexually reproduced varieties. However, long-term clonal reproduction is the main means of propagation in various regions, which leads to the erosion of the species genetic variability and restricts the development of utilization and applications (Wu et al., 2016). Therefore, finding a method that can distinguish different germplasm resources in *P. heterophylla* is urgent.

Previously, the chloroplast genome *rbcL* and *matK* regions, the Internal Transcribed Spacers (ITS) of the nuclear ribosomal DNA, sequence-related amplified polymorphism (SRAP), inter simple sequence repeat (ISSR), and expressed sequence tag-simple sequence repeat (ESR-SSR) have been used to characterize the genetic diversity of *P. heterophylla* germplasm (Yi et al., 2013; Xiao et al., 2014; Xu et al., 2023). Yi et al. (2013) found that the ITS sequences of different *P. heterophylla* varieties had several specific single nucleotide mutation sites and could be used to identify and distinguish samples from nine different producing areas. Xiao et al. (2014) used ISSR to analyze the diversity of 12 *P. heterophylla* cultivars. A total of 73 polymorphic bands were identified, accounting for 89.02% of the total amplified bands, which revealed the clustering of these 12 cultivars into three clades.

With the development of high-throughput sequencing technologies and the decrease in sequencing costs, complete chloroplast genomes assembled from shotgun genomic DNA sequencing provide a more convenient and higher -resolution means to study the relationship among plant cultivated varieties (Straub et al., 2012). The chloroplast genome length is usually between 115 kb and 165 kb, and the length differences are mostly due to inverted repeat (IR) expansion/contraction (Zhu et al., 2016) or gene losses (Lei et al., 2016). As the second-largest plant genome, the chloroplast genome contains rich genetic information for species identification, phylogenetic analysis, and population genetic studies (Palmer, 1991). Dong et al. (2014) employed a chloroplast genomic strategy to design taxon-specific DNA mini-barcodes and applied them to species identification in the *ginsengs*. Liu et al. (2022) obtained chloroplast genome sequences of 24 plant samples in the genus *Atractylodes* and provided a new understanding of their phylogenetic relationship. Utilizing massively parallel sequencing technology for chloroplast genome sequencing in plants can facilitate a better understanding and discrimination of low-level systematic relationships among different taxa in plant classification (Parks et al., 2009). The first *P. heterophylla* chloroplast genome sequence distributed in Korea was reported and indicated that the *P. heterophylla* chloroplast genome has a double-stranded, circular, typically four-segment structure (Kim et al., 2019). However, there is still a lack of population genetic analyses in *P. heterophylla* using chloroplast genomes.

Here, we collected 17 P. *heterophylla* plant samples with remarkable phenotypic characteristics and obtained their chloroplast genome sequences using next-generation sequencing. This study aimed to (1) elucidate the conservation and diversity of *P. heterophylla* chloroplast genomes through comparative genomic approaches; (2) identify the most variable chloroplast genome regions to utilize them as markers for further germplasm conservation and genetic improvement; and (3) determine the relationships between genotypes using the chloroplast genome sequence data.

Materials and methods

Sample collection

In this study, 17 samples of *P. heterophylla* were collected from seven provinces and represented dominant cultivars in China (Table 1). Zheshen No. 1 has an erect growth habit, is unbranched and short, and its leaves are ovate. Its flowers are white, and its roots are spindle -shaped. It is moderately susceptible to leaf spot disease. Zheshen No. 2 has four to six upright branches (more than the Zheshen No. 1), ovate-lanceolate thick leaves, carrot-shaped root tubers, and moderate resistance to leaf spot disease. It does not flower. Zheshen No. 3 is a tetraploid *P. heterophylla* genotype induced by Zheshen No. 1. Zheshen No. 3 has oval, large, thick, dark green leaves, a low seed setting rate, and high-yielding roots. The Minxuan No. 6 and Minxuan No. 7 biotypes have long, oval, and thick leaves, flowering, large root tubers, and are more resistant to viral diseases. The Zherong Datiao was introduced from Guizhou and has characteristically large root tubers. Shitai No. 1 is a variety obtained using a mixed breeding approach. Its plants are upright, tall, and flowering, with round to long oval leaves and long spindle roots. The Guizhou cultivar plants are upright and tall, with oblong-ovate leaves and long spindle roots. The Jurong cultivar is a native cultivated variety with oval and thick leaves and high-yielding roots. The Hunan cultivar plants are upright, tall, and flowering, with long, ovate leaves and fusiform root tubers. The Xuancheng cultivar plants are upright and multi-branched and have tall plants with oblong-ovate leaves and large root tubers. The Shandong cultivar has been domesticated from a wild population. Its plants are tall with branches, and its leaves are oval-lanceolate and thin. The root tuber of the Shandong cultivar is long, spindle-shaped, and thin, and yields for this cultivar are high. The Hebei cultivar was introduced from Shandong and has morphological characteristics like the Shandong cultivar. These samples were identified by Prof. Jingying Chen from the Fujian Academy of Agricultural Sciences.

DNA extraction, library preparation, and high-throughput sequencing

The total genomic DNA from *P. heterophylla* leaf tissues was extracted using a modified CTAB method. DNA quantity and quality were determined using Qubit4.0 (Thermo Fisher Scientific Inc., USA). Subsequently, the genomic DNA was purified and fragmented to construct sequencing libraries (350 bp) using the TruSeq DNA PCR-Free High Throughput Library Prep Kit (Illumina, San Diego, CA). High-throughput sequencing (2 × 150 bp) was performed with the NovaSeq 6000 sequencer (Illumina, San Diego, CA).

Assembly, annotation, and visualization of *P. heterophylla* chloroplast genomes

The PCR-free sequencing data were used to assemble the chloroplast genome sequences of *P. heterophylla* using the GetOrganelle pipeline (Jin et al., 2020). Gene annotation of the

TABLE 1 Collection information of 17 P. *heterophylla* samples.

Sample ID	Cultivar name	Locality
TZ-1	Zheshen No. 1	Yingshan Town, Zherong County, Fujian Province
TZ-2	Zheshen No. 2	Fuxi town, Zherong County, Fujian Province
TZ-3	Zheshen No. 2	Fankeng Town, Fu 'an City, Fujian Province
TZ-4	Zheshen No. 2	Shangbaishi Town, Fu 'an City, Fujian Province
TZ-5	Zheshen No. 3	Yingshan Town, Zherong County, Fujian Province
TZ-6	Minxuan No. 6	Yingshan Town, Zherong County, Fujian Province
TZ-7	Minxuan No. 7	Chuping Town, Zherong County, Fujian Province
TZ-8	Zherong Datiao	Fuxi town, Zherong County, Fujian Province
TZ-9	Shitai No. 1	Niudachang town, Shibing County, Guizhou Province
TZ-10	Guizhou cultivar	Niudachang town, Shibing County, Guizhou Province
TZ-11	Jurong cultivar	Qianxu village, Jurong City, Jiangsu Province
TZ-12	Xuancheng cultivar	Zhongjianshan village, Guangde City, Anhui Province
TZ-13	Xuancheng cultivar	Jinshan Village, Guangde City, Anhui Province
TZ-15	Xuancheng cultivar	Sanhe Village, Xuanzhou District, Anhui Province
TZ-16	Hunan cultivar	Xiaoshajiang Town, Longhui County, Hunan Province
TZ-17	Hebei cultivar	Nanliu Town, Wuji County, Hebei Province
TZ-18	Shandong cultivar	Yushan Town, Linmu County, Shandong Province

chloroplast genome sequences was performed using CpGAVAS2 (Shi et al., 2019) and then manually evaluated and corrected. Graphical maps of *P. heterophylla* chloroplast genome sequences were drawn using OrganellarGenomeDRAW (OGDRAW) (Greiner et al., 2019).

Characterization and comparative analysis of *P. heterophylla* chloroplast genomes

The REPuter (Kurtz et al., 2001) software was used to recognize four types of sequence repeats, including forward (F), reverse (R), complementary (C), and palindromic (P). The minimum repeat size of oligonucleotide repeats was set at 30 bp, and the Hamming distance was set at 3. Tandem repeats were analyzed using the Tandem Repeats Finder (TRF) software (Benson, 1999) with default parameters. Simple sequence repeats (SSRs) were detected using the MIcroSAtellite identification tool (MISA) (Beier et al., 2017). The minimum repeat thresholds of mono-, di-, tri-, tetra-, penta-, and hexanucleotide SSRs were set as 10, 6, 5, 5, 5, and 5, respectively. Primers for SSRs were designed with Primer 3.0 software (Untergasser et al., 2012).

The mVISTA program with the Shuffle-Lagan model (Frazer et al., 2004) was employed to compare the chloroplast genome sequences of *P. heterophylla*. IRscope (Amiryousefi et al., 2018) was used to visualize the contraction and extension of IR boundaries between the four parts of the genome (LSC/IRb/SSC/IRa). Gene rearrangements were observed using the co-linear blocks obtained by the Mauve alignment algorithm (Darling et al., 2004).

ParaAT2.0 software (Zhang et al., 2012) was used to align protein sequences derived from specific protein-encoded DNA sequences extracted from 17 P. *heterophylla* chloroplast genomes. The nucleic acid alignment corresponding to the codon was translated back according to the protein alignment result. KaKs_Calculator 3.0 software (Zhang, 2022) was then used to calculate synonymous (Ks), nonsynonymous (Ka), and Ka/Ks ratios after homologous sequence alignment.

The concatenated protein-coding gene sequences of the 17 *Pseudostellaria* chloroplast genomes were used for phylogenetic analysis, with *Cerastium arvense*, *Gymnocarpos przewalskii*, and *Dianthus caryophyllus* as outgroup species. A maximum likelihood (ML) phylogenetic tree of 1,000 bootstrap replications was constructed using RAxML v8.2.12 (Stamatakis, 2014).

Results

Characterization of *P. heterophylla* chloroplast genomes

The *P. heterophylla* chloroplast genome sequence length ranged from 149,356 bp to 149,592 bp, with a variation of 236 bp among the different samples. Each chloroplast genome had the typical quadripartite structure, with a large single copy (LSC) region (80,994–81,144 bp), a small single copy (SSC) region (16,860 to 17,154 bp), and a pair of IR regions (IRa and IRb) (25,650 to 25,732 bp). The chloroplast genome GC content in all samples ranged from 36.50% to 36.52%, and the GC content in the IR region (approximately 42%) was significantly higher compared to the LSC region and SSC region (approximately 34% and 29%). A total of 111 unique genes were annotated in the *P. heterophylla* chloroplast genomes sequenced, including 77 protein-coding genes, 30 tRNA genes, and four rRNA genes (rrn23S, rrn16S, rrn5S, and rrn4.5S). Among these genes, 46 were related to photosynthesis, and 58 were involved in chloroplast transcription and translation activities. Fifteen genes were in the IR region with two copies, including four protein-coding genes, seven tRNA genes, and four rRNA genes. Seventeen genes contained introns, of which 14 genes (eight protein-coding genes and six tRNA genes) contained one intron, and three genes (*rps12*, *ycf3*, and *clpP*) contained two introns. Small exons were also identified in the *petB*, *petD*, and *rpl16* genes, with lengths of 6 bp, 8 bp, and 9 bp, respectively. In addition, *rps12* was identified as a trans-splicing gene. Further detailed chloroplast genome information is presented in Tables 2, S1 and Figure 1.

Codon usage in *P. heterophylla* chloroplast genomes

The amino acid frequencies, the number of codons, and the relative synonymous codon usage (RSCU) in *P. heterophylla* chloroplast genomes are shown in Table S2. The average RSCU value was 63.97, and the number of codons ranged from 22,012 (TZ-3) to 22,017 (TZ-5). Among the codons, leucine was the amino acid with the most abundant codons. UUU (encoding phenylalanine) and UGC (encoding cysteine) were the most and least used codons, respectively. Almost all amino acids had more than one synonymous codon, except for methionine and tryptophan. Four start codon types were identified in the 77 protein-coding genes. Among them, 73 genes possessed ATG as their start codon, while two genes (*ndhD* and *psbL*) had ACG, one gene (*rps19*) had GTG, and one gene (*ycf1*) had TTG as their start codon. All the samples had the same three stop codon types (TAA, TAG, and TGA). The most used stop codon was TAA (60.98%), followed by TGA (21.95%) and TAG (17.07%).

SSRs, repeat structures, and IRs of *P. heterophylla* chloroplast genomes

For the SSR analysis, 75–84 SSR loci were detected in the *P. heterophylla* chloroplast genomes (Figure 2), among which polyadenine (poly-A) (54.78%, 41–47) and polythymine (poly-T) (38.75%, 29–32) represented the most abundant simple sequence repeats. SSRs and their 500 bp upstream and downstream sequences were extracted, and 69 primer pairs were designed using Primer 3.0 software. After electronic amplification evaluation allowing for two mismatches, four pairs of SSR primers targeting highly polymorphic SSR regions were obtained (Table S3). Sixteen to 21 short tandem repeats were found in the *P. heterophylla* chloroplast genomes (Table S4), ranging in length from 11 to 32 bp, with most located

10.3389/fpls.2023.1163325

TABLE 2 Genes in the chloroplast genome of *P. heterophylla*.

Category	Group	Genes
Miscellaneous group	Acetyl-CoA carboxylase	*accD*
	Cytochrome c biogenesis	*ccsA*
	Maturase	*matK*
Photosynthetic genes	Subunits of ATP synthase	*atpA*, *atpB*, *atpE*, *atpF**, *atpH*, *atpI*
	Chloroplast envelope membrane protein	*cemA*
	ATP-dependentprotease subumitP	*clpP***
	Subunits of NADH dehydrogenase	*ndhA**, *ndhB**, *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *ndhJ*, *ndhK*
	Subumits of cytochrome	*petA*, *petB**, *petD**, *petG*, *petL*, *petN*
	Subunits of photosystem I	*psaA*, *psaB*, *psaC*, *psaI*, *psaJ*
	Subunits of photosystem II	*psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbN*, *psbT*, *psbZ*
	The large subunit of Rubisco	*rbcL*
Transcription and translation-elated genes	Large subunit of ribosome	*rpl14*, *rpl16**, *rpl2*, *rpl20*, *rpl22*, *rpl32*, *rpl33*, *rpl36*
	Small subunit of the ribosome	*rps11*, *rps12***, *rps14*, *rps15*, *rps16**, *rps18*, *rps19*, *rps2*, *rps3*, *rps4*, *rps7*, *rps8*
Protein synthesis and DNA replication	RNA polymerase	*rpoA*, *rpoB*, *rpoC1**, *rpoC2*
RNA genes	Ribosomal RNA genes	*rrn16*, *rrn23*, *rrn4.5*, *rrn5*
	Transfer RNA genes	*trnA-UGC**, *trnC-GCA*, *trnD-GUC*, *trnE-UUC*, *trnF-GAA*, *trnG-GCC*, *trnG-UCC**, *trnH-GUG*, *trnI-CAU*, *trnI-GAU**, *trnK-UUU**, *trnL-CAA*, *trnL-UAA**, *trnL-UAG*, *trnM-CAU*, *trnN-GUU*, *trnP-UGG*, *trnQ-UUG*, *trnR-ACG*, *trnR-UCU*, *trnS-GCU*, *trnS-GGA*, *trnS-UGA*, *trnT-GGU*, *trnT-UGU*, *trnV-GAC*, *trnV-UAC**, *trnW-CCA*, *trnY-GUA*, *trnfM-CAU*
unknown function	Hypothetical chloroplast reading frames(*ycf*)	*ycf1*, *ycf2*, *ycf3***, *ycf4*

*Contains one intron; **Contains two introns.

in the intergenic space (IGS) regions. Twenty-seven to 32 long repeat structures were identified in the *P. heterophylla* chloroplast genomes, including forward, palindromic, reverse, and complement repeats (Table S5). Palindromic was the most common repeat sequence type, accounting for an average of 47.86% of all repeat sequences, followed by forward (40.12%), reverse (11.21%), and complement (0.81%).

The *P. heterophylla* chloroplast genome exhibits four boundaries between the IRs and LSC/SSC regions: LSC-IRb, IRb-SSC, SSC-IRa, and IRa-LSC (Figure S1). The LSC-IRb, IRb-SSC, SSC-IRa, and IRa-LSC boundaries in all samples were located at *rps19*, *ycf1*, *ndhF*, and *rpl2-trnH*, respectively. The *P. heterophylla* chloroplast genomes from the Chinese populations were highly conserved. The nucleotide lengths of *rps19* and *ycf1* located in the IRb region were 195 bp and 105 bp, of *ndhF* located in the IRa region was 56 bp, and of *trnH* from the IRa-LSC boundary was 29 bp.

Candidate markers and Ka/Ks substitution of *P. heterophylla* chloroplast genomes

According to the comparative analysis of the whole chloroplast genome of *P. heterophylla* using the LAGAN program, several regions were variable and were able to distinguish different populations (Figure S2). In terms of genes, the most variable coding genes were *ndhJ*, *ycf1*, and *rpl20*, and the most variable intergenic regions were *psaI-ycf4*, *ycf3-trnS*, *ndhC-trnV*, and *ndhI-ndhG* (Figure 3). Among these genes and intergenic regions, *ycf1* and *ndhI-ndhG* contained a higher number of SNP and MNP polymorphic loci. Particularly, 10 highly polymorphic SNP/MNP loci were identified, which could be used as candidate SNP/MNP markers to distinguish different populations (Table 3). Then, the Mauve algorithm was used to identify the local collinear blocks (LCBs) of the *P. heterophylla* chloroplast genomes, with NC_044183 selected as the reference genome (Figure S3). Among all the chloroplast genomes of the samples, the collinear blocks, including the LSC, SSC, and IR regions, showed relatively high levels of conservation and no gene rearrangements. Thirty-two protein-coding genes with polymorphic sites were used to analyze the synonymous (Ks) and non-synonymous (Ka) substitution rates (Table S6). The average Ka value of the 15 genes was higher than 0.001 (Figure S4), with *rps15*, *rpoC2*, and *rpl20* exhibiting the highest Ka values. Meanwhile, the average Ks value of 17 genes, such as *rps19*, *rps18*, and *rpl14*, was higher than 0.001. The Ka/Ks ratio of all these 32 protein-coding genes ranged from 0.001 to 49.884, with an average value of 19.244. The Ka/Ks ratio of 15 genes was higher than 1, and the gene with the highest Ka/Ks ratio was *rps15* (49.88).

Phylogenetic analysis of *P. heterophylla* chloroplast genomes

To explore the relationships among *P. heterophylla* cultivars, a maximum likelihood (ML) phylogenetic tree was constructed, and *C. arvense*, *G. przewalskii*, and *D. caryophyllus* were selected as out group species (Figure 4). The samples belonging to the Korean *P. heterophylla* population formed a separate cluster from the samples from the Chinese population. In terms of the Chinese *P. heterophylla* population samples, TZ-1, TZ-8, TZ-10–TZ-13,

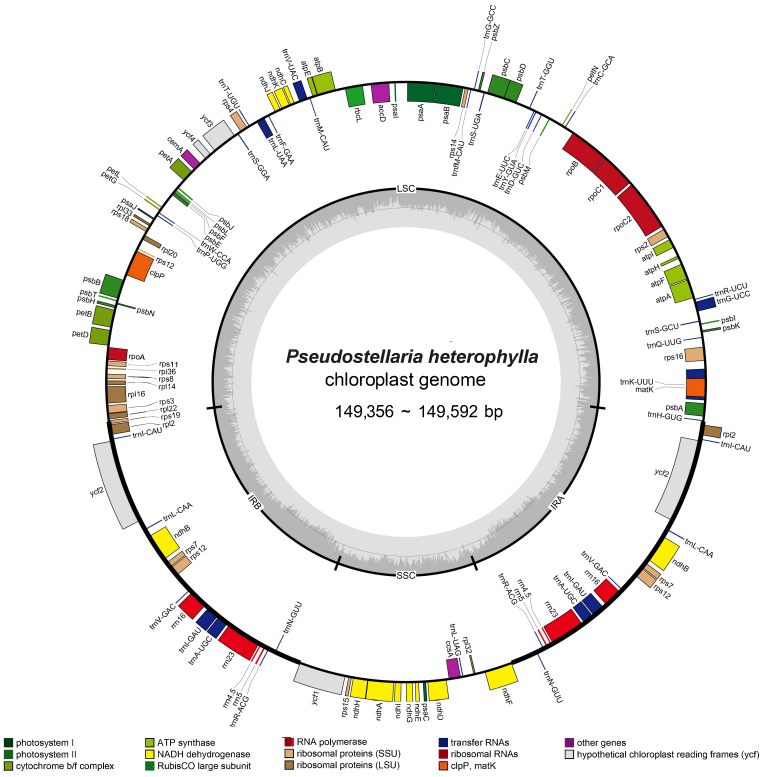


FIGURE 1
Circular chloroplast genome map of *P. heterophylla*. Genes drawn outside the circle are transcribed clockwise, and those inside are counterclockwise. Genes belonging to different functional groups are color-coded. The dark gray area in the inner circle denotes GC content. LSC, large single copy; SSC, small single copy; and IR, inverted repeat.

TZ-15, and TZ-16 were clustered into a big branch, which may be due to the mutual introduction of *P. heterophylla* from Fujian, Jiangsu, Anhui, Hunan, Guizhou, and other locations, resulting in a high similarity of the germplasm resources. TZ-17, TZ-18, and TZ-7 were clustered into a smaller branch that is related to Shandong *P. heterophylla* sources. Three samples of Zheshen No. 2 (TZ-2, TZ-3, and TZ-4) from different places were clustered into a separate branch. TZ-5, TZ-6, and TZ-9 were in separate branches that were located towards the edges of the phylogenetic tree. TZ-6 was selected for its virus resistance. The above results indicated that chloroplast genome sequence analyses could provide useful information for assessing the genetic background of a species. They could be used to assist breeding and provide a molecular – biological basis for cultivar identification.

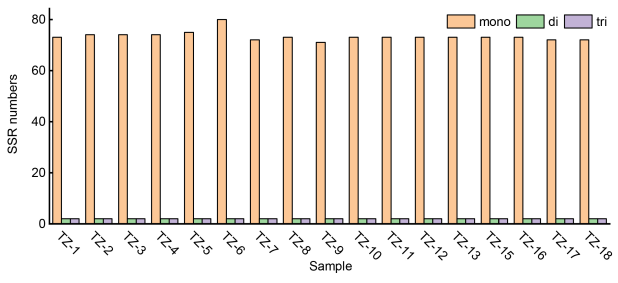


FIGURE 2
Simple sequence repeats (SSRs) in the chloroplast genome of *P. heterophylla*.

Discussion

Distinguishing germplasm resources is essential for plant breeding. Traditional breeding efforts in *P. heterophylla* have usually used plant morphological characteristics, such as leaf size, shape, and thickness; rhizome length, diameter, and texture; plant height; and the number of flowers, to distinguish varieties. However, phenotypes are easily affected by cultivation methods and environmental factors and require long-term observation (Chen et al., 2010). In addition, the irregular introduction of *P. heterophylla* has also impacted the distribution of *P. heterophylla* genetic resources, which affected the uniform collection, classification, and identification of germplasm resources (Xu et al., 2023). *P. heterophylla* cultivation has a history of more than 180 years, and at its earliest stage of cultivation, *P. heterophylla* germplasm resources were mainly derived from wild populations. Since the 1960s, Fujian has successively introduced resources from Jiangsu, Anhui, Zhejiang, Shandong, and other places that formed novel germplasm resources, such as Zheshen No. 1 and Zheshen No. 2. Guizhou Province has no wild *P. heterophylla* populations, and *P. heterophylla* was introduced from Fujian for cultivation in the 1990s. Wild resources of *P. heterophylla* are highly abundant in

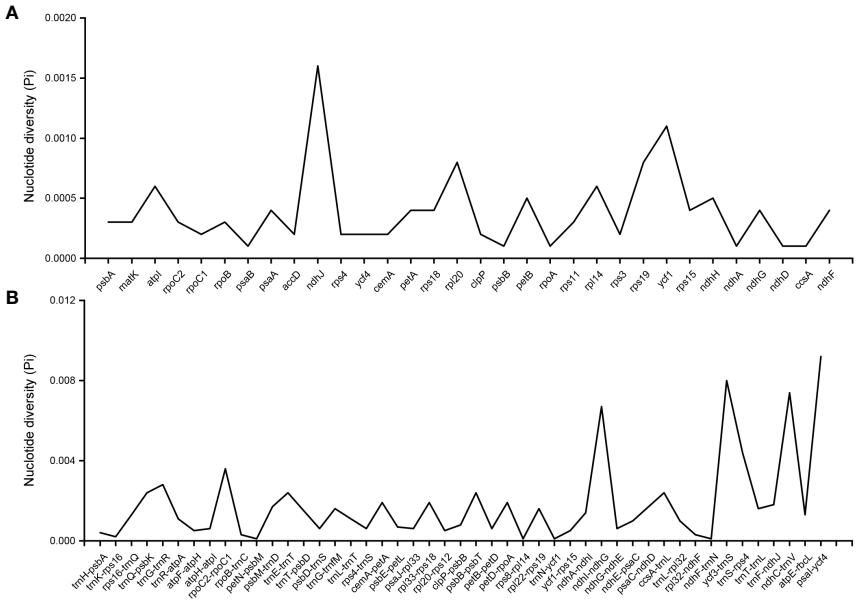


FIGURE 3
The nucleotide diversity of genes and intergenic regions in the *P. heterophylla* chloroplast genomes. **(A)** Coding region; **(B)** Noncoding region.

Jiangsu Province, where there are rarely germplasm introductions from other locations. Due to the use of seeds to raise seedlings, *P. heterophylla* in Jiansu Province is less affected by viral diseases and achieves a higher yield. Interestingly, phylogenetic analysis in this study has provided clues to trace the breeding history of these resources and further verified that the chloroplast genome can provide useful information for analyzing the genetic background of this species. The Korean population was located on the outermost part of the phylogenetic tree as an outgroup. Previous studies reported that *P. heterophylla* is distributed throughout the mountains of Korea, and the morphological characteristics of the Korean population are different from those distributed in China (Choi and Pak, 2000). Therefore, it is possible to obtain new genetic resources or special breeding materials by hybridizing the Korean population with Chinese populations based on the theory of distant hybridization.

Molecular markers derived from the chloroplast genome, such as *rbcL*, *matK*, and *psbA-trnH*, are effective for species identification and phylogenetic resolution (Shi et al., 2011; Liu et al., 2014), and several DNA barcode libraries have been established (Liu et al., 2017). However, species or biotype identification with molecular markers still faces many challenges, especially for closely related species and different populations within species. Previous studies demonstrated that the identification efficiency of DNA barcode

TABLE 3 Candidate polymorphic DNA markers from the chloroplast genome of *P. heterophylla*.

No.	Position*	Polymorphic Type	Variant	Location
1	743–876	SNP	A/T, G/T, T/G	*matK*
2	1,184–2,050	SNP	G/A, T/A, A/T, G/A, G/A	*ndhF*
3	420–900	SNP	A/T, G/C, G/T	*ndhH*
4	2,778–4,100	SNP	A/T, A/G, A/G, A/C	*rpoC2*
5	213–3,605	SNP/MNP	A/G, G/T, A/T, A/C, A/T, A/C, AA/CT, A/T, T/A, T/A, T/A, T/A, T/A, A/T, C/A, A/G, A/T, C/A, G/A, C/T, A/T, A/C	*ycf1*
6	67–247	SNP/MNP	T/A, CAAAATTT/ATTGTAGG, A/T, AA/TT, A/T, T/G	*ndhI_ndhG*
7	21–627	SNP	G/A, A/T, A/T, A/T, A/C, A/T, G/G	*rps16_trnQ-UUG*
8	13–299	SNP	C/T, A/T, A/T	*trnE-UUC_trnT-GGU*
9	7–340	SNP	G/T, C/A, T/A, T/G	*trnL-UAG_rpl32*
10	111–162	SNP	G/A, G/T, C/A	*trnT-GGU_psbC*

*Position is based on the gene and gene spacer alignment data. SNP, single nucleotide polymorphism; MNP, multiple nucleotide polymorphism.

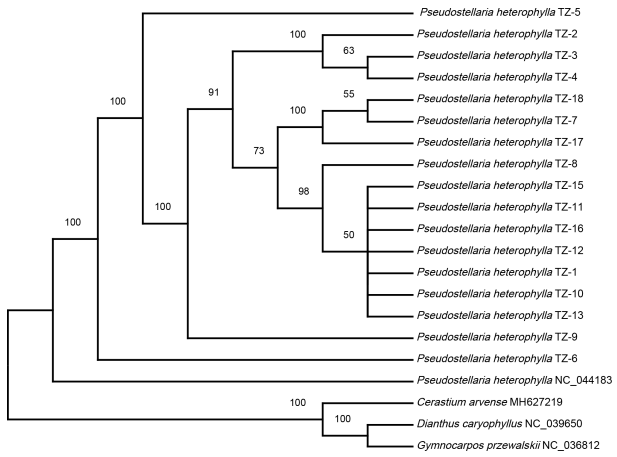


FIGURE 4
Maximum likelihood (ML) phylogenetic tree of *P. heterophylla* inferred from chloroplast genome sequences. Bootstrap values are shown at each node.

markers in specific regions for closely related species was only about 80% (Chen et al., 2014). Several highly informative DNA barcode markers for specific taxa have been developed using comparative analyses of c hloroplast genomes (Zhou et al., 2022). After a comparative analysis of the *Rheum palmatum*, *R. tanguticum*, and *R. officinale* chloroplast genomes, five hypervariable regions (*rps16-trnQ*, *psaA-ycf3*, *psbE-petL*, *ndhF-rpl32*, and *trnT-trnL*) were identified and used as specific DNA barcodes for the identification of 42 samples among *R. tanguticum, R. officinale*, and *R. palmatum* (Li et al., 2022). The *trnl-GAU* intron region was detected to be highly variable and will be useful for future evolutionary studies, although the data from four widely distributed varieties were highly conserved (Wang et al., 2018). The chloroplast genome comparison of *Gentiana* species revealed that the six most InDel-variable loci could be selected as regions for DNA barcode genotyping, confirming that chloroplast genomes could improve the discriminatory capacity for species identification (Zhou et al., 2018). Seven regions (*rpl32-ccsA*, *rpl20-clpP*, *trnC-rpoB*, *ycf1b*, *accD-ycf4*, *ycf1a*, and *psbK-accD*) were identified from the *Pterocarpus* chloroplast genome by quantifying nucleotide diversity and had remarkably higher variability compared to the plant universal barcodes (*rbcL*, *matK*, and *trnH-psbA*) (Jiao et al., 2019). The comparison of the rose chloroplast genome revealed that 15 cpSSRs and 150 flanking single nucleotide variations (SNVs) exhibited higher divergence and stronger power for the genotyping of rose varieties (Li et al., 2020). Moreover, the chloroplast genome can also be used as a super-barcode for phylogenetic and closely related taxon identification studies (Chen et al., 2018).

Conclusion

Using high-throughput sequencing approaches, we obtained the complete chloroplast genome sequences of seventeen *P. heterophylla* varieties. The gene contents and gene orders of the chloroplast genomes were highly conserved. Among these cultivars, 75–84 SSRs, 16–21 short tandem repeats, and 27–32 long repeat structures were detected. Four primer pairs were designed to target highly polymorphic SSR loci. Gene orders were collinear, and IR regions were conserved. Four intergenic regions and three coding genes were found to be highly variable, and ten SNP/MNP sites with polymorphisms were identified and selected for further study. Phylogenetic analysis showed that Chinese populations were clustered into a monophyletic group, in which the non-flowering varieties formed a separate subclade. This study verified that chloroplast genomes could elucidate the relationship among closely related cultivated materials and provide useful information for developing new, highly polymorphic, and informative molecular makers.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: NCBI, PRJNA932041. The GenBank numbers provided are: OQ405025.1~OQ405039.1, OK643505.1, and OK643506.1.

Author contributions

WZ and JC conceived and designed the experiments. WZ and BL performed the experiments. WZ, ZZ, YZ, and YH analyzed and interpreted the data. WZ and ZZ wrote the manuscript. JC revised and approved the manuscript. All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

This study was supported by the Natural Science Foundation of Fujian Province, China (Grant No. 2022J01485), the High-Quality Agricultural Development “5511 “ Collaborative Innovation Project Special Topic of Fujian Provincial People’s Government & Chinese Academy of Agricultural Sciences (Grant No. XTCXGC2021003), the Scientific and Technological Innovation Team of Fujian Academy of Agricultural Sciences (Grant No. CXTD2021014-2), the Construction of “ The Belt and Road “ National Traditional Herbal Medicine Physical Database and Picture Information Database (Grant No. 2018FY100702), the Public Welfare Scientific Research of Fujian Province (Grant No. 2021R1034006), and the Fujian Medicinal Plant Germplasm Resource Nursery (Grant No. ZYBHDWZX202203).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1163325/full#supplementary-material

SUPPLEMENTARY TABLE 1
Genome features of *P. heterophylla* chloroplast genomes.

SUPPLEMENTARY TABLE 2
Codon usage of *P. heterophylla* chloroplast genomes.

SUPPLEMENTARY TABLE 3
Primer pairs for SSRs.

SUPPLEMENTARY TABLE 4
Tandem repeats in *P. heterophylla* chloroplast genomes.

SUPPLEMENTARY TABLE 5
Repeat structures in *P. heterophylla* chloroplast genomes.

SUPPLEMENTARY TABLE 6
Ka/Ks analysis of genes with variable sites.

SUPPLEMENTARY FIGURE 1
The comparison of the *P. heterophylla* chloroplast genome junction boundaries. JLB, junction of LSC and IRb; JLA, junction of LSC and IRb.

SUPPLEMENTARY FIGURE 2
The comparative analysis with LAGAN program of the whole-chloroplast genome of *P. heterophylla*. The x-axis represents the coordinate in the chloroplast genome.

SUPPLEMENTARY FIGURE 3
A comparison of the whole plastid genomes of *P. heterophylla* using the Mauve algorithm. The red LCBs indicate syntenic regions, while the histograms within each block represent the degree of sequence similarity. rRNA, protein-coding, and tRNA gene annotations are denoted by red, white, and green boxes, respectively.

SUPPLEMENTARY FIGURE 4
Ka/Ks analysis of *P. heterophylla* chloroplast gnomes. **(A)** Ka, rate of nonsynonymous substitution; **(B)** Ks, rate of synonymous substitution; **(C)** Ka/Ks, rate of non-synonymous vs. synonymous substitutions.

References

Amiryousefi, A., Hyvönen, J., and Poczai, P. (2018). IRscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics* 34, 3030–3031. doi: 10.1093/bioinformatics/bty220

Beier, S., Thiel, T., Münch, T., Scholz, U., and Mascher, M. (2017). MISA-web: a web server for microsatellite prediction. *Bioinformatics* 33, 2583–2585. doi: 10.1093/bioinformatics/btx198

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi: 10.1093/nar/27.2.573

Chen, S., Pang, X., Song, J., Shi, L., Yao, H., Han, J., et al. (2014). A renaissance in herbal medicine identification: from morphology to DNA. *Biotechnol. Adv.* 32, 1237–1244. doi: 10.1016/j.biotechadv.2014.07.004

Chen, S., Yao, H., Han, J., Liu, C., Song, J., Shi, L., et al. (2010). Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PloS One* 5, 8613. doi: 10.1371/journal.pone.0008613

Chen, X., Zhou, J., Cui, Y., Wang, Y., Duan, B., and Yao, H. (2018). Identification of *Ligularia* herbs using the complete chloroplast genome as a super-barcode. *Front. Pharmacol.* 9. doi: 10.3389/fphar.2018.00695

Choi, K., and Pak, J. H. (2000). A natural hybrid between *Pseudostellaria heterophylla* and *P. palibiniana* (Caryophyllaceae). *Acta Phytotaxon. Geobot.* 50, 161–171. doi: 10.18942/bunruichiri.KJ00001077422

Commission, C. P. (2015). *People's republic of China pharmacopoeia. 2015 Edition* (Beijing, China: China Medical Science and Technology Press).

Darling, A. C., Mau, B., Blattner, F. R., and Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14, 1394–1403. doi: 10.1101/gr.2289704

Dong, W., Liu, H., Xu, C., Zuo, Y., Chen, Z., and Zhou, S. (2014). A chloroplast genomic strategy for designing taxon specific DNA mini-barcodes: a case study on ginsengs. *BMC Genet.* 15, 138. doi: 10.1186/s12863-014-0138-z

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., and Dubchak, I. (2004). VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* 32, W273–W279. doi: 10.1093/nar/gkh458

Greiner, S., Lehwark, P., and Bock, R. (2019). OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* 47, 59–64. doi: 10.1093/nar/gkz238

Jiao, L., Lu, Y., He, T., Li, J., and Yin, Y. (2019). A strategy for developing high-resolution DNA barcodes for species discrimination of wood specimens using the complete chloroplast genome of three *Pterocarpus* species. *Planta* 250, 95–104. doi: 10.1007/s00425-019-03150-1

Jin, J. J., Yu, W. B., Yang, J. B., Song, Y., dePamphilis, C. W., Yi, T. S., et al. (2020). GetOrganelle: a fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. *Genome Biol.* 21, 241. doi: 10.1186/s13059-020-02154-5

Kang, C. Z., Zhou, T., Jiang, W. K., Guo, L. P., Zhang, X. B., Xiao, C. H., et al. (2016). Research on quality regionalization of cultivated *Pseudostellaria heterophylla* based on climate factors. *Zhongguo Zhong Yao Za Zhi* 41, 2386–2390. doi: 10.4268/cjcmm20161303

Kim, Y., Xi, H., and Park, J. (2019). The complete chloroplast genome of prince ginseng, *Pseudostellaria heterophylla* (Miq.) pax (Caryophyllaceae. *Mitochondrial DNA Part B* 4, 2251–2253. doi: 10.1080/23802359.2019.1623127

Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* 29, 4633–4642. doi: 10.1093/nar/29.22.4633

Lei, W., Ni, D., Wang, Y., Shao, J., Wang, X., Yang, D., et al. (2016). Intraspecific and heteroplasmic variations, gene losses and inversions in the chloroplast genome of *Astragalus membranaceus*. *Sci. Rep.* 6, 21669–21669. doi: 10.1038/srep21669

Li, R., Wu, L., Xin, T., Hai, L., Lin, Y. L., Hui, Y., et al. (2022). Analysis of chloroplast genomes and development of specific DNA barcodes for identifying the original species of rhei radix et rhizoma. *Acta Pharm. Sin.* 57, 1495–1505. doi: 10.1038/s41401-021-00781-7

Li, C., Zheng, Y., and Huang, P. (2020). Molecular markers from the chloroplast genome of rose provide a complementary tool for variety discrimination and profiling. *Sci. Rep.* 10, 12188. doi: 10.1038/s41598-020-68092-1

Liu, J., Shi, L., Han, J., Li, G., Lu, H., Hou, J., et al. (2014). Identification of species in the angiosperm family apiaceae using DNA barcodes. *Mol. Ecol. Resour.* 14, 1231–1238. doi: 10.1111/1755-0998.12262

Liu, J., Shi, L., Song, J., Sun, W., Han, J., Liu, X., et al. (2017). BOKP: a DNA barcode reference library for monitoring herbal drugs in the Korean pharmacopeia. *Front. Pharmacol.* 8. doi: 10.3389/fphar.2017.00931

Liu, J., Shi, M., Zhang, Z., Xie, H., Kong, W., Wang, Q., et al (2022). Phylogenomic analyses based on the plastid genome and concatenated nrDNA sequence data reveal cytonuclear discordance in genus Atractylodes (Asteraceae: Carduoideae). *Front. Plant Sci.* 13. doi: 0.3389/fpls.2022.1045423

Ng, T., Liu, F., and Wang, H. (2004). The antioxidant effects of aqueous and organic extracts of *Panax quinquefolium*, *Panax notoginseng*, *Codonopsis pilosula*, *Pseudostellaria heterophylla* and *Glehnia littoralis*. *J. Ethnopharmacol.* 93, 285–288. doi: 10.1016/j.jep.2004.03.040

Palmer, J. D. (1991). "Plastid chromosomes: structure and evolution," in *The molecular biology of plastids*. Eds. L. Bogorad and I. K. Vasil (San Diego: Academic Press), 5–53. doi: 10.1016/B978-0-12-715007-9.50009-8

Parks, M., Cronn, R., and Liston, A. (2009). Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biol.* 7, 84. doi: 10.1186/1741-7007-7-84

Shi, L., Chen, H., Jiang, M., Wang, L., Wu, X., Huang, L., et al. (2019). CPGAVAS2, an integrated plastome sequence annotator and analyzer. *Nucleic Acids Res.* 47, 65–73. doi: 10.1093/nar/gkz345

Shi, L. C., Zhang, J., Han, J. P., Song, J. Y., Yao, H., Zhu, Y. J., et al. (2011). Testing the potential of proposed DNA barcodes for species identification of zingiberaceae. *J. Syst. Evol.* 49, 261–266. doi: 10.1111/j.1759-6831.2011.00133.x

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Straub, S. C. K., Parks, M., Weitemier, K., Fishbein, M., Cronn, R. C., and Liston, A. (2012). Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. *Am. J. Bot.* 99, 349—364. doi: 10.3732/ajb.1100335

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., et al. (2012). Primer3–new capabilities and interfaces. *Nucleic Acids Res.* 40, e115–e115. doi: 10.1093/nar/gks596

Wang, J., Li, C., Yan, C., Zhao, X., and Shan, S. (2018). A comparative analysis of the complete chloroplast genome sequences of four peanut botanical varieties. *PeerJ* 6, e5349. doi: 10.7717/peerj.5349

Wang, Z., Liao, S. G., He, Y., Li, J., Zhong, R. F., He, X., et al. (2013). Protective effects of fractions from *Pseudostellaria heterophylla* against cobalt chloride-induced hypoxic injury in H9c2 cell. *J. Ethnopharmacol.* 147, 540–545. doi: 10.1016/j.jep.2013.03.053

Wu, L., Chen, J., Wu, H., Qin, X., Wang, J., Wu, Y., et al. (2016). Insights into the regulation of rhizosphere bacterial communities by application of bio-organic fertilizer in *Pseudostellaria heterophylla* monoculture regime. *Front. Microbiol.* 7. doi: 10.3389/fmicb.2016.01788

Xiao, C. H., Zhou, T., Jiang, W. K., Ai, Q., Yang, C. G., Xiong, H. X., et al. (2014). Genetic diversity and quality analysis of cultivated. *Pesudostellaria heterophylla.* 45, 1319–1325. doi: 10.7501/j.issn.0253-2670.2014.09.023

Xiao, C., Zhou, T., Jiang, W., Zhao, D., Kang, C., and Liao, M. (2015). Analysis on genetic diversity of phenotypic traits in cultivated *Pseudostellaria heterophylla*. *J. Chin. Med. Mater.* 38, 1600–1606. doi: 10.13863/j.issn.1001-4454.2015.08.009

Xu, L., Li, P., Su, J., Wang, D., Kuang, Y., Ye, Z., et al. (2023). EST-SSR development and genetic diversity in the medicinal plant *Pseudostellaria heterophylla* (Miq.) pax. *J. Appl. Res. Med. Aromat. Plants* 33, 100450. doi: 10.1016/j.jarmap.2022.100450

Yang, Q., Cai, X., Huang, M., and Wang, S. (2020). A specific peptide with immunomodulatory activity from *Pseudostellaria heterophylla* and the action mechanism. *J. Funct. Foods* 68, 103887. doi: 10.1016/j.jff.2020.103887

Yi, J., Liao, F. P., and Zheng, W. W. (2013). Identification of *Pseudostellaria heterophylla* from different idioplasms by analysis of rDNA ITS sequences. *Chin. Tradit. Herb. Drugs* 44, 1318–1322. doi: 10.7501/j.issn.0253-2670.2013.10.022

Zhang, Z. (2022). KaKs_Calculator 3.0: calculating selective pressure on coding and non-coding sequences. *Genomics Proteomics Bioinf.* 20, 536–540. doi: 10.1016/j.gpb.2021.12.002

Zhang, Z., Xiao, J., Wu, J., Zhang, H., Liu, G., Wang, X., et al. (2012). ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. *Biochem. Biophys. Res. Commun.* 419, 779–781. doi: 10.1016/j.bbrc.2012.02.101

Zhou, T., Wang, J., Jia, Y., Li, W., Xu, F., and Wang, X. (2018). Comparative chloroplast genome analyses of species in *Gentiana* section *Cruciata* (Gentianaceae) and the development of authentication markers. *Int. J. Mol. Sci.* 19, 1962. doi: 10.3390/ijms19071962

Zhou, Z., Wang, J., Pu, T., Dong, J., Guan, Q., Qian, J., et al. (2022). Comparative analysis of medicinal plant *Isodon rubescens* and its common adulterants based on chloroplast genome sequencing. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.1036277

Zhu, A., Guo, W., Gupta, S., Fan, W., and Mower, J. P. (2016). Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. *New Phytol.* 209, 1747–1756. doi: 10.1111/nph.13743

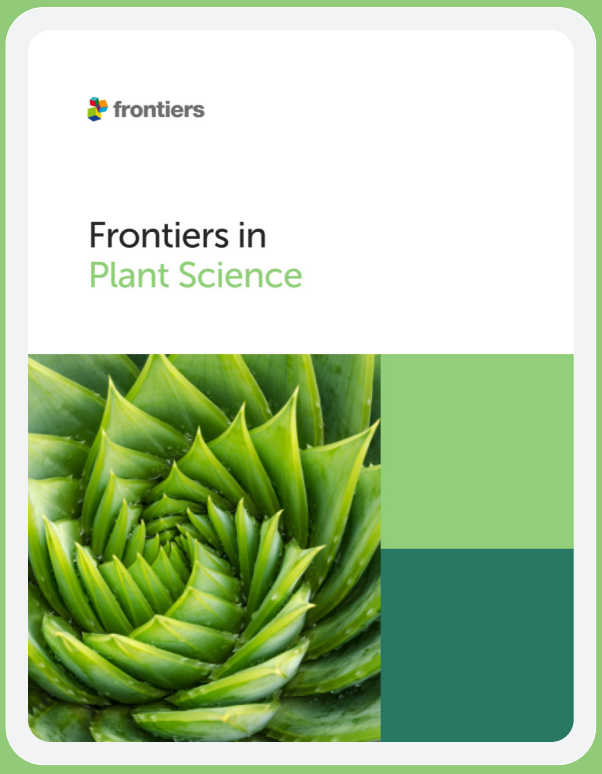
Frontiers in Plant Science

Cultivates the science of plant biology and its applications

The most cited plant science journal, which advances our understanding of plant biology for sustainable food security, functional ecosystems and human health.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact