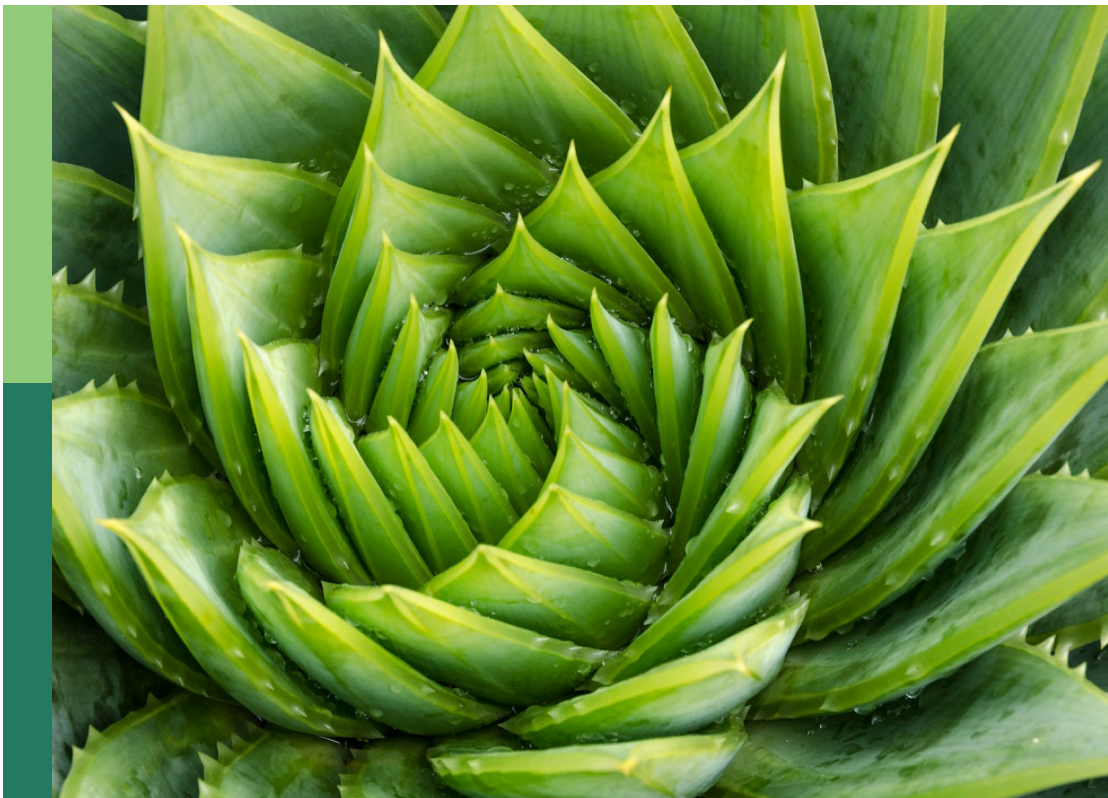
frontiers | Research Topics

Rise to the challenges in plastome phylogenomics

Edited by
Wenpan Dong, Lianming Gao, Peter Poczai, Chao Xu and Yu Song

Published in
Frontiers in Plant Science



ISSN 1664-8714
ISBN 978-2-8325-2724-5
DOI 10.3389/978-2-8325-2724-5

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

Rise to the challenges in plastome phylogenomics

Topic editors

Wenpan Dong — Beijing Forestry University, China

Lianming Gao — Kunming Institute of Botany, Chinese Academy of Sciences (CAS), China

Peter Poczai — University of Helsinki, Finland

Chao Xu — Institute of Botany, Chinese Academy of Sciences (CAS), China

Yu Song — Guangxi Normal University, China

Citation

Dong, W., Gao, L., Poczai, P., Xu, C., Song, Y., eds. (2023). *Rise to the challenges in plastome phylogenomics*. Lausanne: Frontiers Media SA.

doi: 10.3389/978-2-8325-2724-5

Table of contents

frontiers | Frontiers in Plant Science

TYPE Editorial
PUBLISHED 01 June 2023
DOI 10.3389/fpls.2023.1200302

OPEN ACCESS

EDITED AND REVIEWED BY
Jim Leebens-Mack,
University of Georgia, United States

*CORRESPONDENCE
Peter Poczai
peter.poczai@helsinki.fi

RECEIVED 04 April 2023
ACCEPTED 04 May 2023
PUBLISHED 01 June 2023

CITATION
Dong W, Gao L, Xu C, Song Y and Poczai P (2023) Editorial: Rise to the challenges in plastome phylogenomics. *Front. Plant Sci.* 14:1200302. doi: 10.3389/fpls.2023.1200302

Editorial: Rise to the challenges in plastome phylogenomics

Wenpan Dong[1], Lianming Gao[2], Chao Xu[3], Yu Song[4,5] and Peter Poczai[6,7]*

[1]School of Ecology and Nature Conservation, Beijing Forestry University, Beijing, China, [2]CAS Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan, China, [3]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing, China, [4]Key Laboratory of Ecology of Rare and Endangered Species and Environmental Protection (Ministry of Education), Guangxi Normal University, Guilin, Guangxi, China, [5]Guangxi Key Laboratory of Landscape Resources Conservation and Sustainable Utilization in Lijiang River Basin, Guangxi Normal University, Guilin, Guangxi, China, [6]Botany Unit, Finnish Museum of Natural History, University of Helsinki, Helsinki, Finland, [7]Museomics Research Group, Helsinki Institute of Life Science (HiLIFE), Helsinki, Finland

KEYWORDS

plastome phylogenomics, incomplete lineage sorting, introgression, species tree, phylogenetic discordance

Editorial on the Research Topic
Rise to the challenges in plastome phylogenomics

The plastome has a quadripartite organization, encodes about 114 unique genes, and has earned its reputation in plant phylogenomics for ease of obtaining and handling plastome data and its considerable phylogenetic information content (Mehmood et al., 2020). The power of plastome phylogenomics has been exemplified by reconstructing deep to shallow phylogenetic relationships, deducing reticulate evolutionary histories, and phylogenetically placing taxa (Watson et al., 2020; Guo et al., 2023). Despite the upsurge in popularity, decades of studies have raised controversy on the advantages and shortcomings of plastome phylogenomics. Features such as the inheritance characteristics of plastomes, recombination, gene transfer, or specific evolutionary patterns may result in limited or, even worse, incorrect inferences (Gonçalves et al., 2019). Whereas tree discordances are often the primary indication for such problems, the underlying mechanisms should be more extensively explored (Gonçalves et al., 2020; Rose et al., 2021; Doyle, 2022; Kao et al., 2022).

In this Research Topic, we collected studies conducted on solid and thorough plastome phylogenomics throughout the plant tree of life. Chen et al. studied *Isodon* (Schrad. ex Benth.) Spach, a large genus of the Lamiaceae family important for its medicinal properties. Rapid radiation has made it difficult to distinguish species, especially within Clade IV, which contains over 80% of taxa. To elucidate the phylogenetic relationships within the genus, their study used plastome and nrDNA sequences to reconstruct the phylogeny of

approximately 80% of the species. While the results revealed major lineages consistent with previous studies, incongruences were found due to insufficient phylogenetic signal, hybridization, and plastome capture. They revealed that more data from the nuclear genome are needed to resolve relationships within Clade IV and highlighted that nutlet morphology can distinguish the four major clades of *Isodon*.

Xi et al. utilized 19 newly generated plastomes of *Carya* Nutt. (Juglandaceae) species, including the critically endangered species *C. poilanei*, to explore maternal relationships among the subclades of the genus to more comprehensively evaluate plastid genomes, for which variation has not been thoroughly characterized. The results indicated remarkable differences in several plastome features to be highly consistent with the EA-NA disjunction, highlighting the importance of full-length plastomes as an ideal tool for exploring maternal relationships among *Carya* subclades and potentially in other outcrossing perennial woody plants to resolve inter-specifc phylogenetic relationships.

Plastome data were used by Wang et al. to estimate genetic diversity and divergence times, rebuild biogeographic history, and predict potential distribution of the genus *Coptis* Salisb. (Ranunculaceae). With 15 recognized species and high medicinal value, *Coptis* has a conspicuous taxonomy with a unique evolutionary position, distribution pattern, and conservation significance. They revealed that the low nucleotide diversity of *Coptis* plastomes is 0.0067 and the hotspots are located in the *ycf*1 gene. *Coptis* originated in North America and the Japanese archipelago and has a typical Eastern Asian and North American disjunct distribution pattern. The most suitable climatic conditions for *Coptis* were identified, and the study provided insights for future conservation efforts.

Han et al. investigated the genetic structure of *Pistacia chinensis* Bunge (Anacardiaceae), an important tree crop in China known for its high fruit oil content. Through analysis of the plastome and nuclear SNPs of 39 individuals across China, their study identified five clades of *P. chinensis* and the occurrence of hybridization events between highly divergent samples in the subclades. The study suggested that there is much unlocked genetic diversity in this recently domesticated species, which could be exploited for Chinese pistache improvement.

Liu et al. highlighted the complexity of the taxonomic relationships among *Atractylodes* DC (Asteraceae) species, which are cultivated as medicinal herbs in China, Japan, and Korea. The study used high-throughput sequencing to obtain concatenated nuclear ribosomal DNA sequences and plastid genomes from 24 plant samples from five species of *Atractylodes* located in China, of which 23 belonged to members of the *A. lancea* complex. The study identified a mixed clade among this species complex, suggesting the possibility of hybridization or gene introgression.

Zeng et al. analyzed the chloroplast genomes of 123 varieties of mulberry plants from six different species. The study revealed that the analyzed *Morus* taxa should be classified into six species, with two subspecies for *M. alba*. Their research offered valuable insights into the classification, domestication, and breeding improvement of mulberry.

Xiao et al. successfully reconstructed the phylogenetic backbone of *Clematis* L., one of the largest genera of Ranunculaceae, using transcriptome data and plastid genome sequences, nuclear SNP datasets, and single-copy nuclear orthologous genes assembled from genome skimming data. The study found that the assembled datasets could effectively resolve the phylogeny of *Clematis*, showing that rapid species radiation may have caused incomplete lineage sorting, while frequent interspecific hybridization events may have led to cyto-nuclear discordances. The research also provides insights into genome partitioning strategies for future phylogenomic studies of other plant taxa.

Wu et al. focused on the phylogenetic relationships of the family Convolvulaceae, which includes morning glories, bindweeds, and plants of economic importance, such as traditional medicines, ornamentals, and vegetables. They confirmed the monophyly of the family and provided insights into its phylogenomic relationships. The positions of some genera, including *Cuscuta* L. and *Erycibe* Roxb., were uncertain due to variable nucleotide substitution rates. The study also detected numerous plastomic rearrangements, including inversions, duplications, losses of genes, and introns. A rare example of gene transfer from mitochondria to plastids in angiosperms was found in the *Jacquemontia* Choisy plastome.

The taxonomy and phylogenetic relationships of the subgenus *Cerasus*, which includes over 100 species of the genus *Prunus* L. (Rosaceae), are unclear. To address this, Wan et al. reconstructed the phylogenetic tree for 11 species from *P.* subg. *Cerasus* using plastid genome sequences. Phylogenetic analysis revealed the true cherry group to be most similar to the flora of China, with *P. mahaleb* L. forming a distinct subclade. The study provided new insights and potential molecular markers for further research.

The comprehensive evaluation of plastome sequences among plant lineages revealed tree discordance with nuclear genes, elucidating biological and computational factors affecting tree topologies. For phylogenetic studies, a more balanced understanding of the evolutionary history of plastid genes is crucial. The phylogenetic signal of plastid sites and genes varies in support of alternate topologies at various taxonomic levels, as research published in the Research Topic have demonstrated. The disagreement between topologies supported by the greatest number of genes or sites and the topologies supported by the strongest phylogenetic signal may be an indicator of how the phylogenetic signal's variability affects various phylogenetic inference techniques. Complete plastid phylogenies still need to be explored below the order and family levels, despite extensive exploration of this genomic compartment and thorough knowledge of numerous plant phylogenetic nodes. In addition to the common use of plastome data to reconstruct evolutionary connections of different plant groups, we propose comparing plastid phylogenies with nuclear data to uncover possible differences within genomic compartments.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work. All authors of the manuscript have read and agreed to its content and are accountable for all aspect of the accuracy and integrity of the manuscript in accordance with ICMJE criteria.

Acknowledgments

We would like to thank Ana Castro and the Editorial Office of Frontiers in Plant Sciences for their assistance in creating this special collection of studies.

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

References

Doyle, J. J. (2022). Defining coalescent genes: theory meets practice in organelle phylogenomics. *Systematic Biol.* 71, 476–489. doi: 10.1093/sysbio/syab053

Gonçalves, D. J. P., Jansen, R. K., Ruhlman, T. A., and Mandel, J. R. (2020). Under the rug: abandoning persistent misconceptions that obfuscate organelle evolution. *Mol. Phylogenet. Evol.* 151, 106903. doi: 10.1016/j.ympev.2020.106903

Gonçalves, D. J. P., Simpson, B. B., Ortiz, E. M., Shimizu, G. H., and Jansen, R. K. (2019). Incongruence between gene trees and species trees and phylogenetic signal variation in plastid genes. *Mol. Phylogenet. Evol.* 138, 219–232. doi: 10.1016/j.ympev.2019.05.022

Guo, C., Luo, Y., Gao, L. M., Yi, T. S., Li, H. T., Yang, J. B., et al. (2023). Phylogenomics and the flowering plant tree of life. *J. Integr. Plant Biol.* 65, 299–323. doi: 10.1111/jipb.13415

Kao, T. T., Wang, T. H., and Ku, C. (2022). Rampant nuclear-mitochondrial-plastid phylogenomics discordance in globally distributed calcifying microalgae. *New Phytol.* 235, 1394–1408. doi: 10.1111/nph.18219

Mehmood, F., Abdullah,, Ubaid, Z., Shahzadi, I., Ahmed, I., MT, W., et al. (2020). Plastid genomics of nicotiana (Solanaceae): insights into molecular evolution, positive selection and the origin of the maternal genome of Aztec tobacco (*Nicotiana rustica*). *PeerJ* 8, e9552. doi: 10.7717/peerj.9552

Rose, J. P., Toledo, C. A. P., Lemmon, E. M., Lemmon, A. R., and Sytsma, K. (2021). Out of sight, out of mind: widespread nuclear and plastid-nuclear discordance in the flowering plant genus *Polemonium* (Polemoniaceae) suggests widespread historical gene flow despite limited nuclear signal. *Systematic Biol.* 70, 162–180. doi: 10.1093/sysbio/syaa049

Watson, L. E., Siniscalchi, C. M., and Mandel, J. (2020). Phylogenomics of the hyperdiverse daisy tribes: anthemideae, astereae, calenduleae, gnaphalieae, and senecioneae. *J. Systematics Evol.* 58, 841–852. doi: 10.1111/jse.12698

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 08 September 2022
DOI 10.3389/fpls.2022.985488

OPEN ACCESS

EDITED BY
Daniel Pinero,
National Autonomous University of Mexico, Mexico

REVIEWED BY
Ran Wei,
Institute of Botany (CAS), China
Pan Li,
Zhejiang University, China

*CORRESPONDENCE
Lian-Ming Gao
gaolm@mail.kib.ac.cn
Chun-Lei Xiang
xiangchunlei@mail.kib.ac.cn

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

RECEIVED 04 July 2022
ACCEPTED 11 August 2022
PUBLISHED 08 September 2022

CITATION
Chen Y-P, Zhao F, Paton AJ, Sunojkumar P, Gao L-M and Xiang C-L (2022) Plastome sequences fail to resolve shallow level relationships within the rapidly radiated genus *Isodon* (Lamiaceae). *Front. Plant Sci.* 13:985488. doi: 10.3389/fpls.2022.985488

Plastome sequences fail to resolve shallow level relationships within the rapidly radiated genus *Isodon* (Lamiaceae)

Ya-Ping Chen[1], Fei Zhao[1], Alan J. Paton[2], Purayidathkandy Sunojkumar[3], Lian-Ming Gao[1,4]* and Chun-Lei Xiang[1]*

[1]CAS Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, China, [2]Royal Botanic Gardens, Kew, Richmond, United Kingdom, [3]Department of Botany, University of Calicut, Thenhipalam, Kerala, India, [4]Lijiang Forest Biodiversity National Observation and Research Station, Kunming Institute of Botany, Chinese Academy of Sciences, Lijiang, China

As one of the largest genera of Lamiaceae and of great medicinal importance, *Isodon* is also phylogenetically and taxonomically recalcitrant largely ascribed to its recent rapid radiation in the Hengduan Mountains. Previous molecular phylogenetic studies using limited loci have only successfully resolved the backbone topology of the genus, but the interspecific relationships suffered from low resolution, especially within the largest clade (Clade IV) which comprises over 80% species. In this study, we attempted to further elucidate the phylogenetic relationships within *Isodon* especially Clade IV using plastome sequences with a broad taxon sampling of ca. 80% species of the genus. To reduce systematic errors, twelve different plastome data sets (coding and non-coding regions with ambiguously aligned regions and saturated loci removed or not) were employed to reconstruct phylogeny using maximum likelihood and Bayesian inference. Our results revealed largely congruent topologies of the 12 data sets and recovered major lineages of *Isodon* consistent with previous studies, but several incongruences are also found among these data sets and among single plastid loci. Most of the shallow nodes within Clade IV were resolved with high support but extremely short branch lengths in plastid trees, and showed tremendous conflicts with the nrDNA tree, morphology and geographic distribution. These incongruences may largely result from stochasticity (due to insufficient phylogenetic signal) and hybridization and plastid capture. Therefore, the uniparental-inherited plastome sequences are insufficient to disentangle relationships within a genus which has undergone recent rapid diversification. Our findings highlight a need for additional data from nuclear genome to resolve the relationships within Clade IV and more focused studies to assess the influences of multiple processes in the evolutionary

history of *Isodon*. Nevertheless, the morphology of the shape and surface sculpture/indumentum of nutlets is of systematic importance that they can distinguish the four major clades of *Isodon*.

KEYWORDS

genome skimming, Hengduan Mountains, Isodoninae, nutlet, plastid capture

Introduction

Isodon (Schrad. ex Benth.) Spach (Ocimeae, Nepetoideae, Lamiaceae) consists of approximately 100 species, mostly occurring in subtropical to tropical Asia, with two endemic species disjunctly distributed in tropical Africa (Li, 1988; Li and Hedge, 1994; Paton et al., 2009). It is most diverse in southwest China, particularly in the dry valleys of the global biodiversity hotspot Hengduan Mountains (HM), which is considered as the distribution center of the genus (Zhong et al., 2010; Yu et al., 2014; Chen et al., 2019). As a member of the monotypic subtribe Isodoninae established by Zhong et al. (2010), *Isodon* differs from other genera of Ocimeae by the following set of characters: perennial herbs, subshrubs or shrubs, pedunculate and bracteolate cymes, actinomorphic or two-lipped (3/2) calyces, strongly two-lipped (4/1) corollas, and free filaments inserted at the base of the corolla tubes (Wu and Li, 1977; Li, 1988; Paton and Ryding, 1998; Harley et al., 2004). In a worldwide revision of *Isodon*, Li (1988) divided the genus into four sections – sect. *Pyramidium* (Benth.) H.W. Li (7 spp.), sect. *Amethystoides* (Benth.) H.W. Li (7 spp.), sect. *Isodon* (91 spp.), and sect. *Melissoides* (Benth.) H.W. Li (6 spp.) – based on the types of inflorescence, and morphology of fruiting calyx (erect vs. decurved, actinomorphic vs. two-lipped) and corolla tube (saccate vs. gibbous to shortly calcarate on upper side near base). Ten series were further delimited within the largest section (sect. *Isodon*) by Li (1988).

Some species of *Isodon* [e.g., *I. eriocalyx* (Dunn) Kudô, *I. japonicus* (Burm. f.) H. Hara, *I. rubescens* (Hemsl.) H. Hara] have long been used as traditional folk medicine in China and Japan and the genus is abundant in diterpenoids with diverse structural scaffolds and important pharmaceutical functions (Sun et al., 2006; Liu et al., 2017). But despite its apparent value to medicine and understanding of the evolutionary history of the HM flora, several large gaps in our understanding of the taxonomy and systematics of *Isodon* remain and restrict our ability to effectively communicate its infrageneric taxonomic units.

Isodon species are well-known for being difficult to identify, with either diagnostic characters too variable and/or obscure, or having been provided with incomplete descriptions due to a lack of sufficient specimens and/or field investigations (Chen et al., 2019). Low phylogenetic resolution of the infrageneric relationships within *Isodon* from previous molecular phylogenetic studies (Zhong et al., 2010; Yu et al., 2014; Chen et al., 2019) also hampers our ability to assemble a comprehensive and intelligible taxonomy for the genus. Using limited molecular markers, all these studies consistently revealed that Asian *Isodon* can be divided into three strongly supported clades and the relationships within the largest clade which includes more than 80% species of the genus were scarcely resolved with weak to non-existent support. The analyses by Yu et al. (2014) also provided support for a clade of two endemic African species sister to the combination of three Asian clades, whose exact relationship with each other was recovered as equivocal. Moreover, these studies showed that at least three of the aforementioned four sections of *Isodon* proposed by Li (1988) based on the inflorescence types and fruiting calyx and corolla tube morphology were not monophyletic, whereas the color of glands on plants and leaf phyllotaxy might be of systematic significance (Zhong et al., 2010; Yu et al., 2014). The biogeographic study of *Isodon* (Yu et al., 2014) indicated that the genus originated in the Qinghai-Tibetan Plateau (QTP) and adjacent regions in the late Oligocene, and a rapid radiation of the genus triggered by the uplift of QTP and subsequent aridification events might have happened in the late Miocene. They also concluded that the rapid diversification followed by hybridization and introgression might have resulted in the greatest diversity of *Isodon* in HM and the low phylogenetic resolution within the genus (Yu et al., 2014).

Due to the low evolutionary rates, high copy number, uniparental inheritance, lack of recombination or gene duplication compared to the nuclear genome (Birky, 1995; Wicke et al., 2011; Gitzendanner et al., 2018), plastid genomes (plastomes) have been widely used for phylogenetic reconstruction at various levels during the last decade (Yu et al., 2017; Fu et al., 2019; Li et al., 2019, 2021; Xiang et al., 2020; Zhao et al., 2021b; Wang et al., 2022). Meanwhile, an increasing number of studies have demonstrated that substantial improvements in phylogenetic resolution within rapidly radiated genera can be achieved using plastome-scale data sets (e.g., Foster et al., 2018; Ji et al., 2019; Zong et al., 2019; He et al., 2021; Zhao et al., 2021a; Cao et al., 2022). However, most of these infrageneric plastid phylogenomic

studies were based on limited taxon sampling, which may reduce the accuracy of phylogenetic inference (Graybeal, 1998; Hillis, 1998; Zwickl and Hillis, 2002; Heath et al., 2008).

In this study, we carried out the phylogenetic analyses of 86 taxa of *Isodon* using complete plastome and nuclear ribosomal DNA [nrDNA, including 26S, 18S, and 5.8S ribosomal RNA genes, and internal and external transcribed spacers (ITS and ETS)] sequences, representing the most comprehensive taxonomic sampling of the genus to date. We aimed to answer the following questions: (a) whether plastome-scale data can further resolve the shallow-level relationships within *Isodon* based on a comprehensive taxa sampling? (b) if there are any conflicts among different plastome data sets, among single plastid genes, and between the plastid and nrDNA data sets? What might be the causes of these incongruences?

Materials and methods

Taxon sampling

A total of 99 accessions of 86 taxa (including 80 species and six varieties) from the major distribution areas of *Isodon* in East Asia and Africa were sampled as ingroups, representing all recognized sections and series of Li (1988) and all four clades recovered in Yu et al. (2014). Sixteen species from 11 genera of the other six subtribes of Ocimeae were selected as outgroups. Plastomes of 106 individuals were newly generated here from genome skimming data, plus nine plastomes downloaded from GenBank. The sequences of 18S, ITS1, 5.8S, ITS2, and 26S (hereafter referred to as nrITS) were mostly assembled from the genome skimming data, with several sequences obtained by polymerase chain reaction (PCR) amplification or downloaded from GenBank. The ETS sequences generated from the same collections of about half of the species have been published before (Chen et al., 2019), thus these sequences were downloaded from the GenBank. The remaining ETS sequences were newly obtained from PCR amplification. Voucher information and GenBank accession numbers for all sequences are provided in **Supplementary Table 1**. Vouchers of most accessions were deposited at the Herbarium of Kunming Institute of Botany (KUN), Chinese Academy of Sciences.

DNA extraction, amplification, and sequencing

Total genomic DNA was either extracted from silica-gel dried leaves using a modified CTAB method (Doyle and Doyle, 1987) or from herbarium specimens using the DNeasy Plant Mini Kit (Tiangen Biotech, Beijing, China) according to the manufacturer's instructions. Genomic DNA was then sheared into ca. 300 bp fragments, which were used for library construction following standard protocols (NEBNext® Ultra IITMDNA Library Prep Kit for Illumina®). Sequencing with 2 × 150 bp paired-end reads was conducted to generate approximately 2 Gb data for each accession using an Illumina HiSeq 2000 platform (Illumina, San Diego, CA, United States) at BGI Genomics (Shenzhen, Guangdong, China).

PCR mixtures and procedures for the amplification of ITS (including partial 18S, ITS1, 5.8S, ITS2, and partial 26S) and ETS followed those described in Chen et al. (2016a), using primer pairs of 17SE and 26SE (Sun et al., 1994), and ETS-B (Beardsley and Olmstead, 2002) and 18S-IGS (Baldwin and Markos, 1998) for ITS and ETS, respectively. The PCR products were purified and sequenced by the Sangon Biotech (Shanghai, China) on an ABI 3730xl DNA Analyzer (Applied Biosystems, CA, United States).

Plastome and nrDNA assembly and annotation

Adaptors and low-quality reads were removed from the raw data using Trimmomatic v.0.32 (Bolger et al., 2014) with default settings. Subsequently, the *de novo* assembling of clean paired-end reads was carried out using the GetOrganelle Toolkit (Jin et al., 2020), and the resulting contigs were further visualized and edited using Bandage v.0.8.1 (Wick et al., 2015). The newly assembled plastomes were initially annotated with the Plastid Genome Annotator (PGA) (Qu et al., 2019). Using the published plastome of *Isodon amethystoides* (Benth.) H. Hara (GenBank accession number: MT473767; Zhao et al., 2021b) as a reference, the start and stop codons and intron/exon boundaries for protein-coding genes were checked manually in Geneious v.11.0.3 (Kearse et al., 2012). Annotated tRNA genes were verified using the online tRNAscan-SE service (Chan and Lowe, 2019).

The nrITS sequences were *de novo* assembled from the clean data using the GetOrganelle Toolkit (Jin et al., 2020) and the annotation was carried out in Geneious by comparison with the reference sequence of *Perilla frutescens* (L.) Britton (GenBank accession number: KT220698; Cheon et al., 2018). For the sequences of ITS and ETS resulted from Sanger sequencing, trace files with both directions were assembled and edited using Geneious.

Sequence alignment and data set construction

The script "get_annotated_regions_from_gb.py" developed by Zhang et al. (2020) was employed to extract coding and non-coding regions from whole plastomes with one of the inverted repeats (IR) removed. Alignment of individual loci was performed with MAFFT v.7.4.0 (Katoh and Standley, 2013)

using the L-INS-i algorithm and manually adjusted in MEGA 6.0 (Tamura et al., 2013). To minimize the use of loci with limited information, aligned regions less than 25 bp and the conserved rRNAs and tRNAs were excluded from analyses. A total of 207 alignments, including 80 coding genes and 127 non-coding loci, were obtained. The ITS and ETS sequences were aligned separately using the MAFFT plugin in Geneious and then manually adjusted in MEGA.

For the plastome sequences, a total of 12 data sets were generated prior to the phylogenetic reconstruction. Three basic data sets were produced initially: the CR (coding regions; the concatenated 80 coding genes), NCR (non-coding regions; the concatenated 127 non-coding loci), and CR + NCR (the concatenated CR and NCR) data sets. The script "concatenate_fasta.py" developed by Zhang et al. (2020) was used to concatenate the alignments of individual loci. To reduce the impact of misalignment, three data sets CR-GB, CR-GB, and CR + NCR-GB were constructed after using Gblocks v.0.91b (Castresana, 2000; Talavera and Castresana, 2007) to exclude ambiguously aligned regions with default parameters ("Allowed Gap Positions" = "With Half"). Loci with high levels of substitutional saturation were also identified and excluded to construct six data sets. The degree of saturation of all 207 loci was calculated using two indices determined by TreSpEx v.1.1 (Struck, 2014): the slope of the linear regression of plotting patristic distance against uncorrected *p* distances and the R^2 fit of the data to this slope. The higher the slope and R^2, the less saturated is the locus. The slope and R^2 values of all loci are summarized in **Supplementary Table 2**, and the density plots of these values (**Supplementary Figure 1**) were generated using R v.4.0.5 (R Development Core Team, 2021). The threshold of the distribution of slopes for exclusion of loci was determined at the value of 0.85 (**Supplementary Figure 1**). Thus, three data sets (CR-Slope, NCR-Slope, CR + NCR-Slope) were produced by excluding the 72 loci (11 coding and 61 non-coding loci; **Supplementary Table 2**) with slopes lower than 0.85 (determined at which the distribution of the slope values increase significantly). The remaining three data sets (CR-R^2, NCR-R^2, CR + NCR-R^2) were generated after removing the 19 loci (three coding and 16 non-coding loci; **Supplementary Table 2**) located on the left distribution of R^2 (lower than the value of 0.90, at which the distribution of R^2 began to increase; **Supplementary Figure 1**).

For the nuclear sequences, the nrITS and ETS regions were directly concatenated to construct the nrDNA data set.

Phylogenetic analysis

Phylogenetic reconstruction was carried out using maximum likelihood (ML) and Bayesian inference (BI) analyses for each of the 12 plastome data sets and the nrDNA data set. The best partitioning schemes and substitution models for the 12 plastome data sets were estimated by PartitionFinder2 v.2.1.1 (Lanfear et al., 2017) with the selection of all model, RAxML v.8.2.12 (Stamatakis, 2014), the rcluster algorithm (Lanfear et al., 2014), and the corrected Akaike information criterion (AICc). The nrDNA data set was partitioned by the nrITS and ETS region, and the best substitution models for each DNA locus were selected under the AIC using jModelTest 2.1.6 (Darriba et al., 2012).

Partitioned ML analysis was implemented using RAxML-HPC2 v.8.2.12 (Stamatakis, 2014) on XSEDE on the web server Cyberinfrastructure for Phylogenetic Research Science (CIPRES) Gateway (Miller et al., 2010).[1] Analysis of 1,000 rapid bootstrap replicates (-x) was followed by a search for the best-scoring ML tree in a single program run (-f a), both phases using the GTRGAMMA model for nucleotide data and other parameters being the default settings.

Partitioned BI analysis was performed with MrBayes v.3.2.7a (Ronquist et al., 2012) on XSEDE on the CIPRES Gateway. The Markov chain Monte Carlo (MCMC) analyses were run for 20,000,000 generations. Stationarity was considered to be reached when the average standard deviation of split frequencies (ASDSF) fell below 0.01. Trees were sampled at every 1,000 generations with the first 25% discarded as burn-in. The remaining trees were used to build a 50% majority-rule consensus tree.

Gene trees for each of the 207 loci extracted from plastome sequences were also reconstructed using RAxML-HPC2 for 1,000 rapid bootstrap replicates and a search for the best-scoring tree.

Analyses of topological concordance

To explore the concordance/discordance among the 12 data sets of plastome, concatenated gene trees inferred by RAxML from each of the 12 data sets was iteratively used as a reference tree for mapping the remaining trees. The analyses were implemented using PhyParts (Smith et al., 2015) with all trees rooted. To minimize the effect of gene tree estimation error, bipartitions with bootstrap values lower than 70% were ignored for congruence calculations. Results were visualized using the ETE3 Python toolkit (Huerta-Cepas et al., 2016) as implemented in PhyParts_PieCharts.[2] Concordance among single gene trees was also analyzed using PhyParts by mapping the RAxML trees inferred from the 207 loci against the CR + NCR-GB tree, which is the best resolved and most concordant with the remaining 11 concatenated trees. All 207 single gene trees were rooted and only bipartitions with at least 70% BS were considered informative. Finally, incongruence between the CR + NCR-GB tree and the nrDNA tree was visualized using the "tanglegram"

1 http://www.phylo.org/

2 https://github.com/mossmatters/MJPythonNotebooks

function in Dendroscope v.3.8.2 (Huson and Scornavacca, 2012). To minimize spurious disagreement between the two trees due to estimation error, all bipartitions with lower than 50% BS were collapsed.

Morphological and geographical data collection

To test the congruence between morphological and molecular data within *Isodon*, morphological similarities and differences were analyzed based on our previous field investigations and specimen examination. Specimens of *Isodon* and related genera from 29 herbaria (A, AU, BM, CDBI, CSFI, E, G, GXMI, HHBG, HIB, IBK, IBSC, K, KUN, KYO, L, LBG, LE, MW, NAS, P, PE, S, SYS, SZ, TAI, TI, W, and WUK; abbreviations follow Thiers, 2022) were examined. The systematic significance of nutlet morphology was also investigated using light microscopy (LM) and scanning electron microscopy (SEM). Mature nutlets were collected from natural populations and herbarium collections. A total of 61 taxa of *Isodon* were selected (**Supplementary Table 4**). Except for some species where only 1–5 nutlets were available, most species were represented by 10–20 mature nutlets. Measurements and LM analysis were carried out using the Keyence VHX-6000 digital microscope (Keyence Corporation, Osaka, Japan). For SEM, nutlets were directly mounted onto stubs and sputter-coated with gold. Micromorphological observations were conducted using the Hitachi S4800 (Hitachi Ltd., Tokyo, Japan) or Zeiss EVO LS10 (Carl Zeiss NTS, Oberkochen, Germany) scanning electron microscopes at 10 kV. Terminologies used for nutlet description followed those of Budantsev and Lobova (1997) and Moon et al. (2009).

To test the congruence between geographical and molecular data within *Isodon*, the extant distribution data of each species was compiled from taxonomic and floristic literature (Wu and Li, 1977; Li, 1988; Murata and Yamazaki, 1993; Suddee et al., 2004; Paton et al., 2009) and herbarium records. Five biogeographic regions were delimited based on Yu et al. (2014): (A) QTP and adjacent regions; (B) tropical Asia; (C) eastern Asia; (D) Japan; (E) tropical Africa.

Results

Plastome features of *Isodon*

All *Isodon* plastomes newly assembled here displayed the typical quadripartite structure, including a large single copy (LSC) region and a small single copy (SSC) region separated by two inverted repeat (IRa and IRb) regions (**Supplementary Figure 2**). The complete plastomes ranged in length from 151,923 bp [*I. yuennanensis* (Hand.-Mazz.) H. Hara] to 152,824 bp [*I. ternifolius* (D. Don) Kudô] (**Supplementary Figure 2** and **Supplementary Table 3**). The LSC regions ranged from 82,906 bp (*I. yuennanensis*) to 83,640 bp (*I. ternifolius*), the SSC regions varied between 17,532 bp [*I. megathyrsus* (Diels) H. Hara] and 17,729 bp [*I. lophanthoides* var. *graciliflorus* (Benth.) H. Hara], and the IR regions ranged from 25,671 bp [*I. phyllopodus* (Diels) Kudô] to 25,759 bp [*I. rugosus* (Wall. ex Benth.) Codd]. The overall GC content of the *Isodon* plastomes was similar and ranged from 37.5% in *I. ramosissimus* (Hook. f.) Codd to 37.7% in *I. pharicus* (Prain) Murata and *I. adenanthus* (Diels) Kudô (**Supplementary Table 3**). Plastomes of *Isodon* consist of a total of 114 unique genes, including 80 protein coding genes, 30 tRNA genes, and four rRNA genes, and the gene order was highly conserved (**Supplementary Figure 2** and **Supplementary Table 3**).

Phylogenetic relationships and conflicts among data sets

The sequence characteristics of all 12 plastome data sets and the nrDNA data set were summarized in **Table 1**. The alignment length of the nrDNA matrix was 6,371 bp, but with the highest percentage of parsimonious-informative sites (9.65%). Among the 12 plastome data sets, the alignment length of the CR + NCR data set (129,827 bp) was the longest, and that of the NCR-Slope data set was the shortest (28,319 bp), whereas the NCR-GB data set contained the highest percentage of variable sites (15.94%) and parsimonious-informative sites (7.82%).

All 12 plastome data sets yielded almost identical tree topologies at deep nodes and exhibited largely congruent relationships at shallow nodes with some differences or conflicts (**Supplementary Figure 3**). For the three data sets consisting of coding and/or non-coding regions and without further treatment, the phylogenetic tree resulted from the complete plastome (CR + NCR data set) had the highest resolution, followed by the NCR and the CR data sets. Several hard incongruences can be found between the NCR tree and the CR tree, whereas these conflicting nodes in the CR + NCR tree were either collapsed into polytomies or similar to the better supported ones. After removing sites that were poorly aligned, the phylogenetic relationships and support values in the CR + NCR-GB tree and the CR-GB tree were almost unchanged (even slightly better) compared with the tree topologies prior to the treatment. However, the resolution of the NCR-GB tree was slightly poorer than that of the NCR tree, as significantly larger number of misaligned sites pruned from the NCR data set than from the CR data set (**Table 1**). After excluding possible saturated loci with R^2 values lower than 0.90 (three coding and 16 non-coding loci), the resolution of the CR-R^2 tree was largely consistent with that of CR tree, but were slightly poorer in the NCR-R^2 tree and CR + NCR-R^2 tree compared with the NCR tree and CR + NCR tree, respectively. Likewise, the CR-slope

TABLE 1 Detailed characteristics of all plastome data sets and the combined nrDNA data set used in present study.

Data set	No. of loci	No. of sites [bp]	No. of variable sites [bp]	No. of parsimonious-informative sites [bp]
CR	80	69,430	6,696 (9.64%)	3,357 (4.84%)
CR-GB	80	68,694	6,649 (9.68%)	3,339 (4.86%)
CR-Slope	69	64,327	6,275 (9.75%)	3,115 (4.84%)
CR-R^2	77	67,762	6,561 (9.68%)	3,282 (4.84%)
NCR	127	60,397	8,833 (14.62%)	4,223 (6.99%)
NCR-GB	127	50,540	8,054 (15.94%)	3,950 (7.82%)
NCR-Slope	66	28,319	3,020 (10.66%)	1,432 (5.06%)
NCR-R^2	111	55,486	8,077 (14.56%)	3,862 (6.96%)
CR + NCR	207	129,827	15,529 (11.96%)	7,580 (5.84%)
CR + NCR-GB	207	119,234	14,703 (12.33%)	7,289 (6.11%)
CR + NCR-Slope	135	92,646	9,295 (10.03%)	4,547 (4.91%)
CR + NCR-R^2	188	123,248	14,638 (11.88%)	7,144 (5.80%)
nrDNA	2	6,371	936 (14.69%)	615 (9.65%)

tree was also consistent with the CR tree after removing eleven coding loci with slope values lower than 0.85. In contrast, with nearly half (61 out of 127) of the non-coding loci excluded, the NCR-slope tree had the lowest resolution among the 12 concatenated trees. Interestingly, the support value of the sister clade of *Isodon* formed by ten outgroup species was improved by ambiguously aligned regions and saturated loci being removed.

The ML tree resulted from the CR + NCR-GB data set was used as the main reference tree (Figure 1), as it was the best resolved and most concordant with the remaining 11 concatenated trees. The BI tree was also consistent with the ML tree of the same data set and among the 12 data sets, therefore, only PP support values of the BI tree were superimposed together with BS support values on the nodes of the ML tree (Figure 1). *Isodon* was shown to be monophyletic (BS = 100%/PP = 1.00) and four major clades (Clades I–IV) were recovered within the genus. As the first diverging clade, Clade I (BS = 100%/PP = 1.00) consisted of 11 taxa which can be further divided into two strongly supported subclades. The first subclade comprised *I. lophanthoides* (Buch.-Ham. ex D. Don) H. Hara var. *lophanthoides*, *I. lophanthoides* var. *graciliflorus*, *I. atroruber* R.A. Clement, *I. villosus* Y.P. Chen and H. Peng, and *I. scrophularioides* (Wall. ex Benth.) Murata, and the other subclade included *I. oreophilus* (Diels) A.J. Paton and Ryding, *I. flavidus* (Hand.-Mazz.) H. Hara. *I. phyllopodus*, and *I. yuennanensis*. The next split lineage was composed of the two African endemic species *I. ramosissimus* and *I. schimperi* (Vatke) J.K. Morton (Clade II; BS = 100%/PP = 1.00). Clade III consists of the two individuals of *I. ternifolius* (BS = 100%/PP = 1.00), which was further sister to the largest Clade IV (BS = 100%/PP = 1.00). Despite a few shallow nodes were collapsed into polytomies or weakly supported, most of the nodes within Clade IV were resolved with robust support. However, the internal branch lengths within Clade IV were extremely short. The single gene trees were largely unresolved, with some of the loci exhibiting conflicts with the concatenated data set (Supplementary Figure 4).

Consistent with the plastid tree, the nrDNA tree (Supplementary Figure 5) also recovered *Isodon* as a monophyletic group (BS = 87%/PP = 1.00) consisting of four major lineages: Clade I (BS = 96%/PP = 1.00), Clade II (BS = 100%/PP = 1.00), Clade III (BS = 100%/PP = 1.00), and Clade IV (BS = 72%/PP = 1.00). Even though the relationships within Clade IV of the nrDNA tree were almost unresolved with a large polytomy, the plastid-nuclear tanglegram demonstrated widespread discordances (Figure 2). In contrast with the plastid tree, Clade II was shown to be sister to Clade I in the nrDNA tree (BS = 89%/PP = 1.00). The placement of *I. scrophularioides* within Clade I also varied between the two genomic data sets that it was sister to the subclade formed by *I. yuennanensis*, *I. flavidus*, *I. phyllopodus*, and *I. oreophilus* in the nuclear tree (BS = 61%/PP = 0.87; Supplementary Figure 5). Most of the incongruences were located within Clade IV, which comprised over 80% sampled taxa of *Isodon*.

Nutlet morphology of *Isodon* species

Nutlets of *Isodon* species are 0.9–2.1 mm long, 0.6–1.3 mm wide, and range from light yellow or yellowish brown to dark brown (Figure 3 and Supplementary Table 4). All species are characterized by ovate or rounded nutlets (length/width $\leq$ 1.8) having rounded apex (opposite to the areole area), except for *I. ternifolius*, which has trigonous-oblong nutlets (length/width $\geq$ 2) with acute apex. Six different surface types of nutlets can be recognized within *Isodon*, including psilate, reticulate, reticulate-papillate, striate, cellular, and glandular (and pubescent) types (Figure 4 and Supplementary Table 4). Both the psilate and reticulate types can be found in

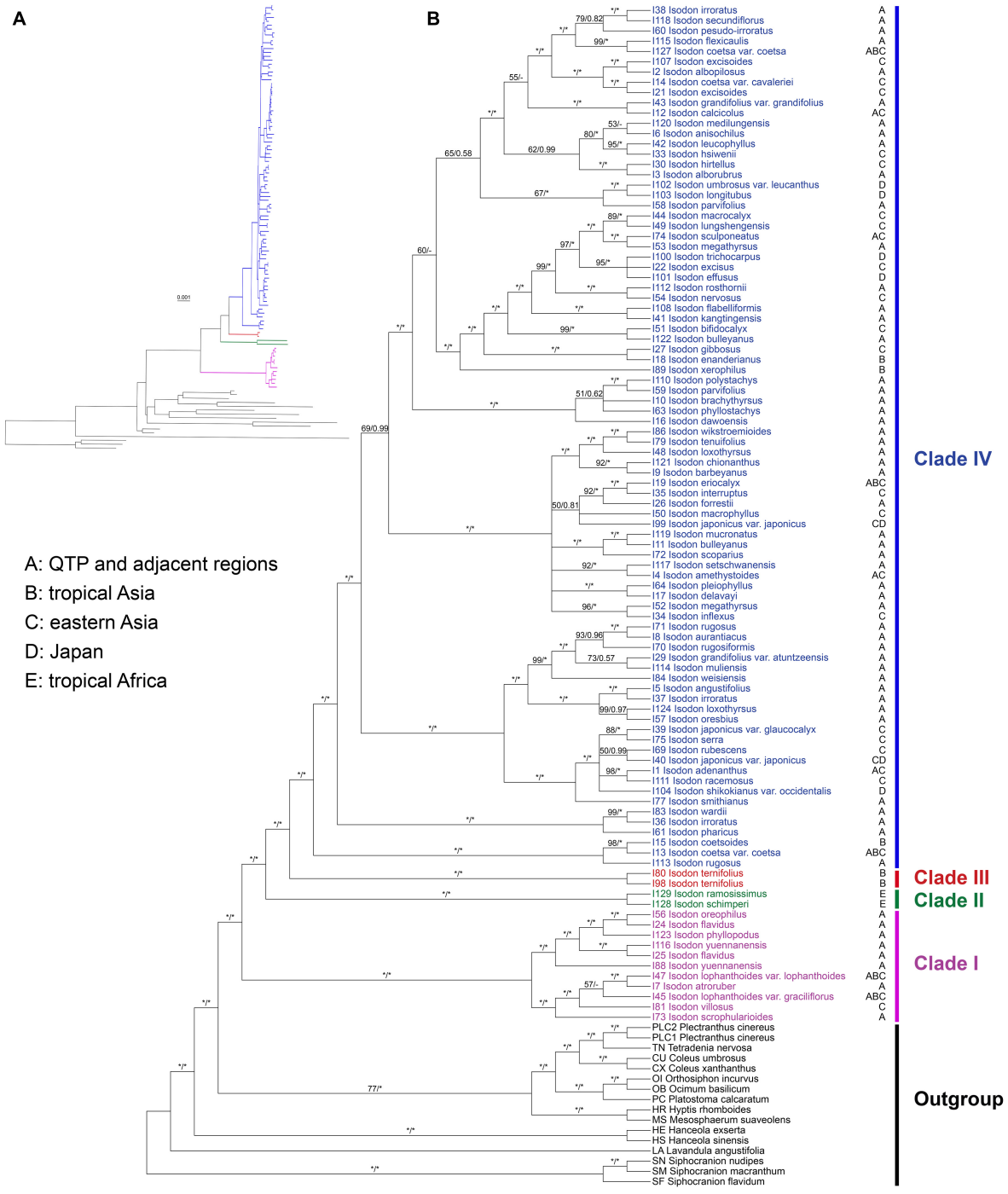


FIGURE 1
Cladogram **(A)** and phylogram **(B)** of the maximum-likelihood tree of *Isodon* derived from the plastid phylogenomic analysis of a concatenated data set including 80 coding and 127 non-coding loci with ambiguously aligned regions removed (CR + NCR-GB data set). Support values ≥ 50% BS or 0.50 PP are displayed above the branches ("*" indicates a support value = 100% BS or 1.00 PP, "–" indicates a support value < 0.50 PP). The distribution area(s) of each *Isodon* species is shown beside the tip and letters coding for areas follow Yu et al. (2014).

species of Clade I, and the reticulate-papillate type is confined to the two African species of Clade II. *Isodon ternifolius* (Clade III) has distinct nutlets with a striate surface, whereas most species of Clade IV possess nutlets with cellular surfaces. Nutlets with glandular or glandular and pubescent trichomes are only possessed by the remaining species of Clade IV.

FIGURE 2
Tanglegram comparing plastid (maximum-likelihood tree inferred from the CR + NCR-GB data set) and nuclear (maximum-likelihood tree inferred from the nrDNA data set) trees, optimized in Dendroscope to minimize line crossings. All clades with < 50% BS have been collapsed.

Discussion

Phylogenetic relationship within *Isodon*

This examination of infrageneric relationships for *Isodon* was aimed at resolving the phylogenetic ambiguities found by prior phylogenetic studies. Our study has increased taxonomic sampling to include all four sections and ten series, and provided additional geographic sampling to accommodate the distribution of *Isodon*, especially from HM. Our results recover four monophyletic clades (Clades I–IV) of *Isodon* species similar to Yu et al. (2014). However, the combination of plastome-scale data and nrDNA sequences of *Isodon* with our extensive field investigation and morphological studies enables us to further discuss the relationships within each clade.

Relationships within Clade I

All species recovered in Clade I are perennial herbs with reddish-brown glands all over the plants, the latter character

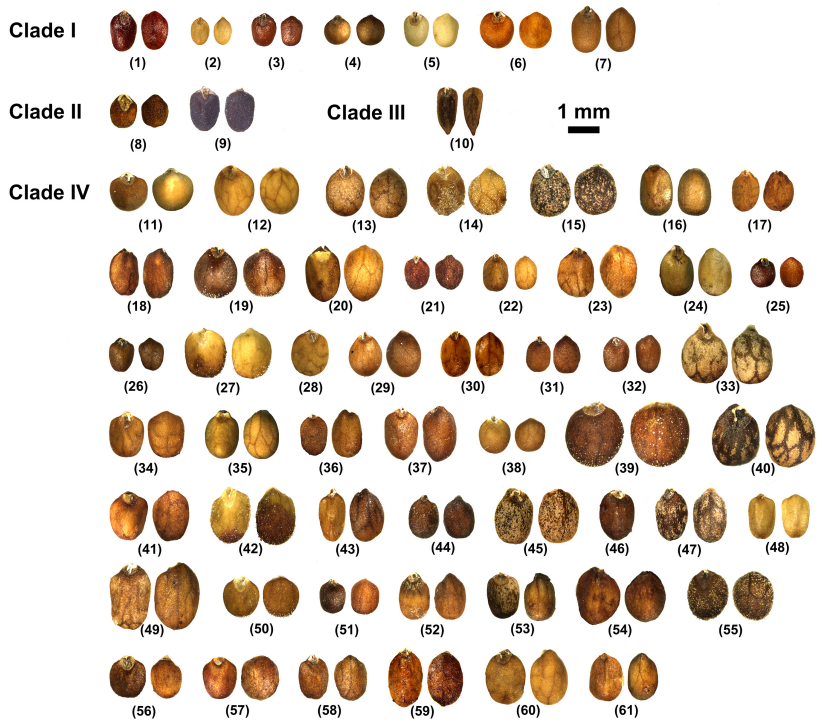


FIGURE 3
Nutlet morphology of taxa of *Isodon*. (1) *I. atroruber*; (2) *I. lophanthoides* var. *lophanthoides*; (3) *I. lophanthoides* var. *graciliflorus*; (4) *I. oreopilus*; (5) *I. phyllopodus*; (6) *I. scrophularioides*; (7) *I. villosus*; (8) *I. ramosissimus*; (9) *I. schimperi*; (10) *I. ternifolius*; (11) *I. adenanthus*; (12) *I. albopilosus*; (13) *I. alborubrus*; (14) *I. amethystoides*; (15) *I. angustifolius*; (16) *I. anisochilus*; (17) *I. aurantiacus*; (18) *I. barbeyanus*; (19) *I. bifidocalyx*; (20) *I. bulleyanus*; (21) *I. coetsa* var. *coetsa*; (22) *I. coetsa* var. *cavaleriei*; (23) *I. dawoensis*; (24) *I. delavayi*; (25) *I. enanderianus*; (26) *I. eriocalyx*; (27) *I. excisus*; (28) *I. forrestii*; (29) *I. gibbosus*; (30) *I. grandifolius* var. *atuntzeensis*; (31) *I. hirtellus*; (32) *I. interruptus*; (33) *I. irroratus*; (34) *I. japonicus* var. *glaucocalyx*; (35) *I. kangtingensis*; (36) *I. leucophyllus*; (37) *I. loxothyrsus*; (38) *I. lungshengensis*; (39) *I. macrocalyx*; (40) *I. macrophyllus*; (41) *I. megathyrsus*; (42) *I. nervosus*; (43) *I. oresbius*; (44) *I. parvifolius*; (45) *I. pharicus*; (46) *I. phyllostachys*; (47) *I. pleiophyllus*; (48) *I. polystachys*; (49) *I. pseudoirroratus*; (50) *I. rosthornii*; (51) *I. rubescens*; (52) *I. rugosus*; (53) *I. scoparius*; (54) *I. sculponeatus*; (55) *I. serra*; (56) *I. setchwanensis*; (57) *I. smithianus*; (58) *I. ternuifolius*; (59) *I. wardii*; (60) *I. weisiensis*; (61) *I. wikstroemioides*.

was considered a useful synapomorphy of this clade by Zhong et al. (2010) and Yu et al. (2014), as opposed to the colorless glands in the remaining species of *Isodon* (**Supplementary Figure 6**). A third kind of glands was discovered in our recently published new species, *I. aurantiacus* Y.P. Chen and C.L. Xiang, from Tibet, China (Chen et al., 2017), which is embedded within Clade IV (**Figure 1** and **Supplementary Figure 5**). The plants of this species are covered with orange glands (**Supplementary Figure 6**).

Two subclades can be further recognized within Clade I, whereas the placement of *I. scrophularioides* exhibits the largest discrepancy that it is recovered within different subclades in the plastid and nrDNA trees. The nutlet morphology seems to support the position of *scrophularioides* in the nrDNA tree. Our examination of the exine ornamentation on nutlets shows a close relationship between *I. oreophilus* and *I. phyllopodus* with *I. scrophularioides*, as their surfaces are reticulate, unlike the psilate surface sculpture in species (*I. atroruber* and *I. lophanthoides*) belonging to the other subclade of Clade I (**Figure 4** and **Supplementary Table 4**). However, a more comprehensive sampling of species with reddish-brown glands is needed to verify the hypothesis.

Species of Clade I are most diverse in the Himalayas and tropical Asia. Based on our current understanding of the taxonomy of *Isodon*, a total of 20 species have been reported with reddish-brown glands (Li, 1988; Zhong et al., 2010; Chen et al., 2016b; Kumar et al., 2016; Ranjan et al., 2022). The continuous discovery of new species possessing this character during recent years possibly indicates an underestimation of the species diversity of Clade I.

Synapomorphy of Clade II

The systematic placement of the two species of *Isodon* endemic to Africa, *I. ramosissimus* and *I. schimperi*, was first elucidated by Yu et al. (2014), using nrITS, three plastid markers, and a low-copy nuclear gene. Our results here are congruent with that of Yu et al. (2014) that the two species were recovered in a monophyletic clade either sister to the large clade formed

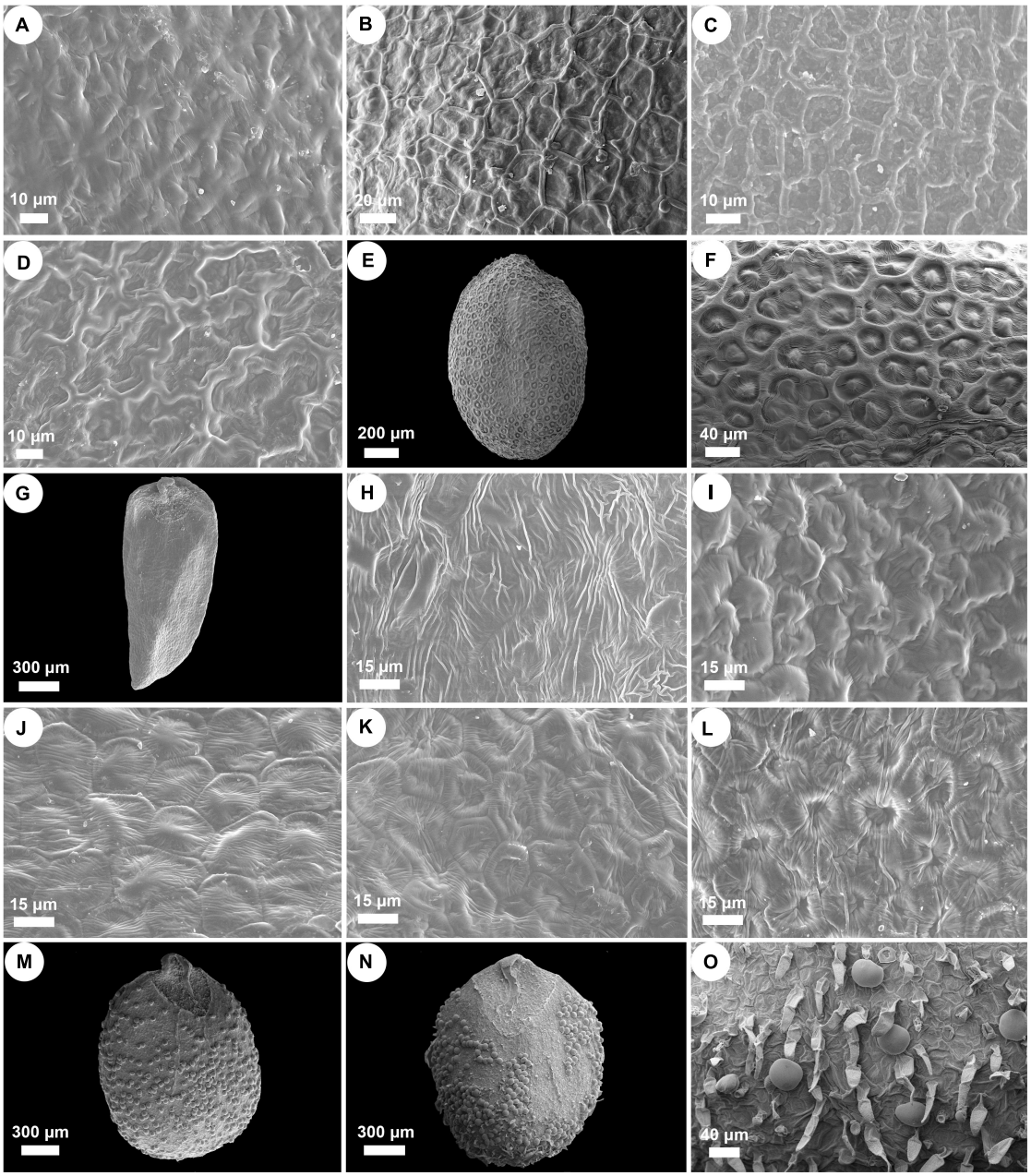


FIGURE 4
SEM micrographs of nutlet surface pattern in *Isodon*. **(A)** Psilate surface. *I. lophanthoides*. **(B–D)** Reticulate surface. **(B)** *I. oreophilus*; **(C)** *I. phyllopodus*; **(D)** *I. scrophularioides*. **(E,F)** Reticulate-papillate surface. *I. schimperi*. **(G,H)** Striate surface. *I. ternifolius*. **(I–L)** Cellular surface. **(I)** *I. barbeyanus*; **(J)** *I. dawoensis*; **(K)** *I. scoparius*; **(L)** *I. smithianus*. **(M–O)** Glandular or glandular and pubescent surface. **(M)** *I. amethystoides*; **(N,O)** *I. serra*.

by Clade III and Clade IV in the plastid tree (Figure 1) or Clade I in the nuclear tree (Supplementary Figure 5). Based on molecular phylogenetic analyses, chromosome evolution, and biogeographic study, Yu et al. (2014) inferred that the two African species were of hybrid origin between members of Clade I and Clades III–IV, both of which likely formed through allopolyploidy; an overland migration from Asia to Africa through Arabia in early to middle Miocene and the opening of the Red Sea in the middle Miocene probably resulted in the present disjunct distribution of *Isodon*.

Though the evolutionary history of Clade II has been elucidated by previous study, the synapomorphy of this clade was unclear. Both species are herbs with loose panicles, straight corolla tubes, and long exserted stamens, morphologically

similar to *I. lophanthoides* and related species from Clade I but with colorless glands on the plants. Our morphological study of nutlets further support the distinctiveness of the two endemic species from Africa that they possess reticulate-papillate nutlet surfaces differing from all other species of *Isodon*, which can be regarded as a synapomorphy of this clade (**Figure 4**).

Synapomorphy of Clade III

By including two accessions of *I. ternifolius* in the analysis, we provide additional evidence that supports *I. ternifolius* as a distinct lineage within *Isodon* (Zhong et al., 2010; Yu et al., 2014). *Isodon ternifolius*, which is widely distributed from tropical China to South and Southeast Asia (Li, 1988; Suddee et al., 2004), is reported by Zhong et al. (2010) to be distinguished by a 3–4-whorled leaf phyllotaxy. However, opposite phyllotaxy can be observed on some individuals of a whorled-leaf population of *I. ternifolius*, and sometimes even on an individual with whorled-leaves (Li, 1988; Suddee et al., 2004; Y.P. Chen, pers. obs.). Furthermore, whorled leaves can occasionally be found in other species of *Isodon* (Y.P. Chen, pers. obs.). Invariable transitions between phyllotaxy such as these indicate that the current diagnosis of *I. ternifolius* cannot be properly applied, and must be adjusted to a more stable synapomorphic character. Our examination of nutlets (**Figures 3**, **4**) suggest that irrespective of phyllotaxy, *I. ternifolius* has trigonous-oblong nutlets with an acute apex, unlike the nutlets of other *Isodon* species, which are ovate and have an obtuse or rounded apex. Moreover, the striate nutlet ornamentation of *I. ternifolius* is also unique in the genus.

Relationships within Clade IV

The majority (ca. 85%) of *Isodon* species in the phylogenetic trees are recovered in Clade IV. Despite that the relationships within this clade are much better resolved in the plastid tree (**Figure 1**), compared with that of nrDNA tree which primarily includes a large polytomy and scant substructures associated with short branch lengths (**Supplementary Figure 5**), strong cytoplasmic-nuclear discordance is also revealed (**Figure 2**). Moreover, some nodes within Clade IV are still poorly supported or show conflicts among the 12 concatenated trees, indicating that whole plastome sequence is not a panacea in resolving phylogenetic relationships within rapidly radiated genera.

As the most morphologically diverse group of *Isodon*, species of Clade IV are herbs or shrubs with a variety of leaves, inflorescences, and trichomes. It has been difficult to find out the synapomorphy for such a clade. Interestingly, herbal species are mostly distributed in eastern Asia and Japan, whereas the shrubby species, which account for a large proportion of Clade IV, are mainly accustomed to the dry valleys in HM. Two types of nutlet surface are restricted to Clade IV as revealed by our morphological study, with most species sharing the cellular nutlet ornamentation (**Figure 4**). The remaining species are characterized with glandular or glandular and pubescent nutlets; all these species [except for *I. tenuifolius* (W.W. Sm.) Kudô] are herbs from the forest area in eastern Asia and Japan.

Causes of phylogenetic incongruences

The 12 concatenated plastome data sets resulted in largely congruent and better resolved topologies (**Figure 1** and **Supplementary Figure 3**) than that of previous studies. However, it can be noted that the branch lengths within Clade IV are extremely short and some nodes receive weak support values or exhibit conflicts among these data sets, especially among different sequence types (coding vs. non-coding) (**Supplementary Figure 3**). Moreover, concordance analyses of single gene trees demonstrated that most of the plastid genes are uninformative for majority of nodes, but strongly supported conflict at some nodes is also recovered (**Supplementary Figure 4**). Systematic (e.g., modeling error, misalignment, saturation, and limited taxon sampling) and stochastic errors (e.g., limited phylogenetic signal/information) as well as biological factors (e.g., selection and possibility of genes with distinct evolutionary histories, incomplete lineage sorting) might potentially contribute to these observed discordances (Walker et al., 2019). Based on a comprehensive sampling and our efforts to prune regions with poor alignment and exclude loci with high levels of substitutional saturation, systematic errors fail to explain most of the conflicts as the resulting data sets recovered largely concordant infrageneric relationships (**Supplementary Figure 3**). Previous studies (Walker et al., 2019; Zhang et al., 2020) suggested that stochasticity could be a primary source of conflict among plastid loci. Phylogenetic relationships inferred from genes with limited information (usually with short length or slow-evolving) could be spurious and misleading. For a genus which has undergone recent rapid radiation, most single plastid regions harbor too few variable sites to be informative and thus unable to accurately infer relationships within *Isodon*. Although potential sources of conflict such as selection and heteroplasmy (i.e., distinct plastomes coexist within a single organism resulted from plastid recombination, transfer, and/or complex mutational dynamics) have been documented from many other taxa (Lee-Yaw et al., 2019; Ramsey and Mandel, 2019), they still warrant further investigation in *Isodon*.

From our analyses, it is clear that well-supported tree obtained from plastome data is in strong conflict with the tree inferred from nuclear data (**Figure 2**). Moreover, the infrageneric relationships recovered within the plastid tree are also incongruent with the current taxonomy (mainly based on

morphology and ecology) and species distribution. Patterns of variation in nuclear markers often agree with morphologically defined species boundaries (e.g., Rautenberg et al., 2010; Schuster et al., 2018; Ogishima et al., 2019). However, the sister relationship of *I. interruptus* (C.Y. Wu and H.W. Li) H. Hara and *I. brachythyrsus* (C.Y. Wu and H.W. Li) H. Hara which receives high support (BS = 96%/PP = 1.00) in the nrDNA tree is in contrast with their significant differences in the morphology of inflorescence (spike-like panicles with bracts longer than cymes in *I. interruptus* vs. loose panicles with bracts shorter than cymes in *I. brachythyrsus*), calyx (tubular campanulate vs. broadly campanulate), and corolla (yellow with tube slightly saccate abaxially near base vs. white to pale lavender with tube calcarate abaxially near base), whereas *I. interruptus* shares all these features with *I. muliensis* (W.W. Sm.) Kudô (Chen and Xiang, 2019). The minor differences between *I. interruptus* and *I. muliensis* are that *I. interruptus* has shorter habits and petioles and smaller leaves. Despite the constraints of nrDNA sequences (e.g., multiple rDNA arrays, concerted evolution, pseudogenes, and secondary structure) (Álvarez and Wendel, 2003; Feliner and Rossello, 2007), the limited substructures recovered within Clade IV in the nuclear tree are mostly formed by species with similar morphology and from the same distribution area. In contrast, similar patterns of morphology and geographic distribution are hardly found in the plastid tree (**Figure 1**). For example, the two sister groups recovered in the nrDNA tree (**Supplementary Figure 5**), *I. gibbosus* (C.Y. Wu and H.W. Li) H. Hara and *I. lungshengensis* (C.Y. Wu and H.W. Li) H. Hara, *I. xerophilus* (C.Y. Wu and H.W. Li) H. Hara and *I. enanderianus* (Hand.-Mazz.) H.W. Li, can also be supported by their shared morphology, habitat, and distribution. Plants of *I. gibbosus* and *I. lungshengensis* are perennial herbs usually grow at moist streamside in the bordering area of Guangxi, Hunan, and Guizhou Provinces, China, while *I. enanderianus* and *I. xerophilus* are shrubs restricted to the dry valleys in southern Yunnan Province (Wu and Li, 1977; Li and Hedge, 1994). However, in the plastid tree, *I. gibbosus* falls sister to *I. enanderianus*, both of which are distantly related to *I. lungshengensis* and *I. xerophilus* (**Figure 1**).

Discordances between nuclear tree/morphological taxonomy and plastid tree are much often attributed to interspecific hybridization, which could be the gene flow in the nuclear genome or plastid capture. The discordant placement of the two African species (Clade II) and *I. scrophularioides* likely indicates a hybrid origin, as has been shown in Yu et al. (2014). Except for these species, several natural hybrids have been described from Japan (Murata and Yamazaki, 1993); our recent field investigations have also detected some putative hybrids in sympatric populations of *Isodon* that will require further study (Chen, 2017; Y.P. Chen, pers. obs.). Although genetic exchange within the genus seems to be frequent, both chromosome incompatibilities and/or backcrossing with one parent could erased most signals of gene flow in the nuclear genome, but these signals may have been retained in the plastid genome by chance or by selection. Moreover, in seed plants with maternally inherited organelles, interspecific exchange of plastid loci occurs more frequently than that of nuclear genes (Rieseberg and Soltis, 1991). Therefore, the majority of topological conflicts between the plastid and nuclear trees possibly is a consequence of plastid capture, i.e., the plastid genome of one species is replaced by an alien one through intrageneric hybridization and introgression as well as recurrent backcrossing, while the nuclear genome remains largely unchanged (Rieseberg and Soltis, 1991; Tsitrone et al., 2003; Chan and Levin, 2005; Stegemann et al., 2012). Molecular phylogenetic studies of Japanese *Isodon* also revealed that most populations belonging to the same species are rarely monophyletic in the plastid tree, but group together in the nuclear tree, indicating that plastid capture may have frequently occurred between *Isodon* species from Japan (Maki et al., 2010; Ogishima et al., 2019). Previous studies frequently recovered geographic patterns of large-scale structuring in plastid haplotypes (e.g., Rautenberg et al., 2010; Pham et al., 2017; Schuster et al., 2018). However, this geographic congruence with infrageneric relationships in the plastid tree is neither supported by our result nor by the analyses of Japanese *Isodon* (Maki et al., 2010; Ogishima et al., 2019). Most species forming monophyletic clades/sister groups in the plastid tree do not show sympatric distribution or geographic proximity (**Figure 1**), suggesting the plastid capture might be a consequence of ancient (rather than ongoing) interspecific hybridization within shared areas in the past.

Taken as a whole, our results suggest that stochasticity stemming from recent rapid radiation and limited phylogenetic signal, widespread hybridization and plastid capture, independently or in concert, may contribute to a complex evolutionary history of *Isodon*. Moreover, our results demonstrate that plastome data does not accurately reflect species relationships and is insufficient to resolve shallow level relationships within a rapidly radiated genus like *Isodon*. To achieve for a better resolved phylogeny of the genus, significant larger data sets derived from nuclear genome with next-generation sequencing technology might be helpful, such as targeted enrichment (Gardner et al., 2020; Reichelt et al., 2021; Giaretta et al., 2022), deep genome skimming (Liu et al., 2021), and RNA-seq (Nevado et al., 2016; Jin et al., 2021; Kong et al., 2022; Xia et al., 2022), which are compatible with more comprehensive tools for disambiguating genetic lineages and causes of gene tree conflicts.

Conclusion

Isodon, as revealed by plastid and nuclear phylogenetic data, is comprised of four clades, the fourth (Clade IV) of which consists of the largest portion of *Isodon* (85% of species).

The tree topology at deep nodes largely corresponds with previous molecular phylogenetic studies of *Isodon*, whereas we further reveal that Clade II which consisting of the two African endemic species differs from the remaining *Isodon* species by the reticulate-papillate nutlet surfaces. Clade III is composed of individuals of *I. ternifolius* with both whorled and opposite phyllotaxy, while our morphological examination demonstrates that nutlets with acute and conspicuously trigonous apex is exclusive to all individuals of *I. ternifolius*. Phylogenetic relationships within Clade IV are much better resolved in the plastid tree than that in previous studies, but highly incongruent with the nrDNA tree and morphology and distribution. Species of Clade IV have cellular nutlet surface type and the distinct nutlets with trichomes, the latter type is restricted to species mostly from eastern Asia and Japan. Our results also suggest that multiple processes might have been involved in the evolutionary history of *Isodon* and plastome sequences fail to further resolve the shallow-level relationships within the genus. Large-scale molecular markers from the nuclear genome might contribute to a better understanding of the infrageneric relationships of the rapidly radiated *Isodon*.

Data availability statement

The original contributions presented in this study are publicly available. The raw sequence data were deposited in the NCBI BioProject database under the accession number: PRJNA863331, and the sequence alignments and resulting tree files were deposited in figshare at https://doi.org/10.6084/m9.figshare.20170727.

Author contributions

Y-PC, L-MG, and C-LX conceived this study. Y-PC and FZ analyzed the data. Y-PC drafted the manuscript with contributions from other authors. All authors collected the materials, read, and approved the final manuscript.

Funding

This work was supported by the Yunnan Fundamental Research Projects (Grant Nos. 202101AU070067 and 202101AT070159) to Y-PC, the Large-Scale Scientific Facilities of the Chinese Academy of Sciences (Grant No. 2017-LSFGBOWS-02) and the Key Basic Research program of Yunnan Province, China (Grant No. 202101BC070003) to L-MG, the Ten Thousand Talents Program of Yunnan (Grant No. YNWR-QNBJ-2018-279), the Yunnan Fundamental Research Projects (Grant No. 2019FI009), and the open research project of the Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences to C-LX, and the Science and Engineering Research Board of the Government of India (Grant No. CRG/2018/003499) to PS.

Acknowledgments

We would like to thank the staff of following herbaria for their kind assistance in research facilities: BM, CDBI, E, IBK, IBSC, K, KUN, KYO, LE, MW, NAS, PE, SZ, TI. We are grateful to Ya-Huang Luo for his help in data analyses, and to Zhi-Jia Gu for his technical assistance in SEM. Thanks are also extended to En-De Liu, Hong-Jin Dong, Lei Jiang, and staff of the Germplasm Bank of Wild Species in Southwest China for their help in collection of specimens and leaf material, and to the reviewers for their valuable suggestions that greatly improved our 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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.985488/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Locus-specific saturation indices for 207 loci shown as density plots (distribution). **(A)** R^2 of the linear regression between patristic and uncorrected pairwise distances. **(B)** Slopes of the linear regression between patristic and uncorrected pairwise distances. Dashed line indicates starting shoulder value. Red regions on the left-hand side contain loci that might be saturated.

SUPPLEMENTARY FIGURE 2
Gene map of the complete plastomes of *Isodon* species. Genes inside and outside of the circle are transcribed in the clockwise and counterclockwise directions, respectively. Genes belonging to different functional categories are color-coded. LSC, large single copy; SSC, small single copy; IR, inverted repeat.

SUPPLEMENTARY FIGURE 3
Cladogram of the maximum-likelihood tree estimated from the CR + NCR-GB data set used as a reference to evaluate conflict and concordance among the trees estimate from the remaining 11 data sets. Pie charts depict conflict amongst the input trees, with the blue, green, red, and gray slices representing the proportion of input bipartitions concordant, conflicting (supporting a single main alternative topology), conflicting (supporting various alternative topologies), and uninformative (< 70% BS) at each node in the CR + NCR-GB tree, respectively. The numbers above and below each branch are the number of bipartitions concordant and conflicting with that particular node, respectively.

SUPPLEMENTARY FIGURE 4
Summary of gene tree conflict against the reference CR + NCR-GB tree. Pie charts depict conflict amongst the input locus trees, with the blue, green, red, and gray slices representing the proportion of the 207 trees concordant, conflicting (supporting a single main alternative topology), conflicting (supporting various alternative topologies), and uninformative (< 70% BS) at each node in the PCN tree, respectively. The numbers above and below each branch are the number of bipartitions concordant and conflicting with that particular node, respectively.

SUPPLEMENTARY FIGURE 5
Phylogram of the maximum-likelihood tree of *Isodon* inferred from the nrDNA data set. Support values ≥ 50% BS or 0.50 PP are displayed above the branches ("*" indicates a support value = 100% BS or 1.00 PP, "–" indicates a support value < 0.50 PP).

SUPPLEMENTARY FIGURE 6
Glands with different colors on leaves and calyces of *Isodon* species. **(A,B)** reddish-brown, *I. lophanthoides* var. *lophanthoides*; **(C,D)** colorless, I. serra; **(E,F)** orange, *I. aurantiacus*.

References

Álvarez, I., and Wendel, J. F. (2003). Ribosomal ITS sequences and plant phylogenetic inference. *Mol. Phylogenet. Evol.* 29, 417–434. doi: 10.1016/S1055-7903(03)00208-2

Baldwin, B. G., and Markos, S. (1998). Phylogenetic utility of the external transcribed spacer (ETS) of 18S-26S rDNA: Congruence of ETS and ITS trees of Calycadenia (Compositae). *Mol. Phylogenet. Evol.* 10, 449–463. doi: 10.1006/mpev.1998.0545

Beardsley, P. M., and Olmstead, R. G. (2002). Redefining Phrymaceae: The placement of Mimulus, tribe Mimuleae, and Phryma. *Am. J. Bot.* 89, 1093–1102. doi: 10.3732/ajb.89.7.1093

Birky, C. W. (1995). Uniparental inheritance of mitochondrial and plastid genes: Mechanisms and evolution. *Proc. Natl. Acad. Sci. U.S.A.* 92, 11331–11338.doi: 10.1073/pnas.92.25.11331

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

Budantsev, A. L., and Lobova, T. A. (1997). Fruit morphology, anatomy and taxonomy of tribe Nepeteae (Labiatae). *Edinb. J. Bot.* 54, 183–216. doi: 10.1007/s10265-006-0023-6

Cao, Q., Gao, Q. B., Ma, X. L., Zhang, F. Q., Xing, R., Chi, X. F., et al. (2022). Plastome structure, phylogenomics and evolution of plastid genes in Swertia (Gentianaceae) in the Qinghai-Tibetan Plateau. *BMC Plant Biol.* 22:195. doi: 10.1186/s12870-022-03577-x

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17, 540–552. doi: 10.1093/oxfordjournals.molbev.a026334

Chan, K. M. A., and Levin, S. A. (2005). Leaky prezygotic isolation and porous genomes: Rapid introgression of maternally inherited DNA. *Evolution* 59, 720–729. doi: 10.1111/j.0014-3820.2005.tb01748.x

Chan, P. P., and Lowe, T. M. (2019). tRNAscan-SE: Searching for tRNA genes in genomic sequences. *Methods Mol. Biol.* 1962, 1–14.doi: 10.1007/978-1-4939-9173-0_1

Chen, Y. P., Hu, G. X., Zhao, F., Peng, H., and Xiang, C. L. (2017). Taxonomic notes on Isodon (Lamiaceae) in China, II: I. aurantiacus, a new species from Tibet, China. *Ann. Bot. Fenn.* 54, 239–243. doi: 10.5735/085.054.0606

Chen, Y. P. (2017). *Taxonomic and molecular phylogenetic studies on Isodon (Schrad. ex Benth.) Spach (Lamiaceae) in China. [dissertation]*. Beijing: University of Chinese Academy of Sciences.

Chen, Y. P., and Xiang, C. L. (2019). Taxonomic notes on Isodon (Nepetoideae, Lamiaceae) in China, III: Resurrection of I. brachythyrsus, I. chionanthus, and I. kangtingensis. *Nord. J. Bot.* 37, 1–13. doi: 10.1111/njb.02240

Chen, Y. P., Xiang, C. L., Sunojkumar, P., and Peng, H. (2016b). Isodon villosus (Nepetoideae, Lamiaceae), a new species from Guangxi, China. *Phytotaxa* 268, 271–278. doi: 10.11646/phytotaxa.268.4.5

Chen, Y. P., Drew, B. T., Li, B., Soltis, D. E., Soltis, P. S., and Xiang, C. L. (2016a). Resolving the phylogenetic position of Ombrocharis (Lamiaceae), with reference to the molecular phylogeny of tribe Elsholtzieae. *Taxon* 65, 123–136. doi: 10.12705/651.8

Chen, Y. P., Wilson, T. C., Zhou, Y. D., Wang, Z. H., Liu, E. D., Peng, H., et al. (2019). Isodon hsiwenii (Lamiaceae: Nepetoideae), a new species from Yunnan, China. *Syst. Bot.* 44, 913–922. doi: 10.1600/036364419X15710776741486

Cheon, K. S., Jeong, I. S., Kim, K. H., Lee, M. H., Lee, T. H., Lee, J. H., et al. (2018). Comparative SNP analysis of chloroplast genomes and 45S nrDNAs reveals genetic diversity of Perilla species. *Plant Breed. Biotech.* 6, 125–139. doi: 10.9787/PBB.2018.6.2.125

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2: More models, new heuristics and parallel computing. *Nat. Methods* 9:772. doi: 10.1038/nmeth.2109

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.

Feliner, G. N., and Rossello, J. A. (2007). Better the devil you know? Guidelines for insightful utilization of nrDNA ITS in species level evolutionary studies in plants. *Mol. Phylogenet. Evol.* 44, 911–919. doi: 10.1016/j.ympev.2007.01.013

Foster, C. S. P., Henwood, M. J., and Ho, S. Y. W. (2018). Plastome sequences and exploration of tree-space help to resolve the phylogeny of riceflowers (Thymelaeaceae: Pimelea). *Mol. Phylogenet. Evol.* 127, 156–167. doi: 10.1016/j.ympev.2018.05.018

Fu, C. N., Mo, Z. Q., Yang, J. B., Ge, X. J., Li, D. Z., Xiang, Q. Y., et al. (2019). Plastid phylogenomics and biogeographic analysis support a Laurasian origin and rapid early radiation of Cornales in the Mid-Cretaceous. *Mol. Phylogenet. Evol.* 140:106601. doi: 10.1016/j.ympev.2019.106601

Gardner, E. M., Johnson, M. G., Pereira, J. T., Puad, A. S. A., Arifiani, D., Wickett, N. J., et al. (2020). Paralogs and off-target sequences improve phylogenetic resolution in a densely-sampled study of the breadfruit genus (Artocarpus, Moraceae). *Syst. Biol.* 70, 558–575. doi: 10.1093/sysbio/syaa073

Giaretta, A., Murphy, B., Maurin, O., Mazine, F. F., Sano, P., and Lucas, E. (2022). Phylogenetic relationships within the hyper-diverse genus Eugenia (Myrtaceae: Myrteae) based on target enrichment sequencing. *Front. Plant Sci.* 12:759460. doi: 10.3389/fpls.2021.75946

Gitzendanner, M. A., Soltis, P. S., Yi, T. S., Li, D. Z., and Soltis, D. E. (2018). "Plastome phylogenetics: 30 years of inferences into plant evolution," in *Advances in botanical research*, eds S. M. Chaw and R. K. Jansen (London: Academic Press), 293–313. doi: 10.1016/bs.abr.2017.11.016

Graybeal, A. (1998). Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst. Biol.* 47, 9–17. doi: 10.1080/106351598260996

Harley, R. M., Atkins, S., Budantsev, A. L., Cantino, P. D., Conn, B. J., Grayer, R., et al. (2004). "Labiatae," in *The families and genera of vascular plants*, eds K. Kubitzki and J. W. Kadereit (Berlin: Springer), 167–275.

He, J., Lyu, R., Luo, Y., Lin, L., Yao, M., Xiao, J., et al. (2021). An updated phylogenetic and biogeographic analysis based on genome skimming data reveals convergent evolution of shrubby habit in clematis in the pliocene and pleistocene. *Mol. Phylogenet. Evol.* 164:107259. doi: 10.1016/j.ympev.2021.107259

Heath, T. A., Hedtke, S. M., and Hillis, D. M. (2008). Taxon sampling and the accuracy of phylogenetic analyses. *J. Syst. Evol.* 46, 239–257.

Hillis, D. M. (1998). Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Syst. Biol.* 47, 3–8. doi: 10.1080/106351598260987

Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* 33, 1635–1638. doi: 10.1093/molbev/msw046

Huson, D. H., and Scornavacca, C. (2012). Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. *Syst. Biol.* 61, 1061–1067. doi: 10.1093/sysbio/sys062

Ji, Y. H., Yang, L. F., Chase, M. W., Liu, C. K., Yang, Z. Y., Yang, J., et al. (2019). Plastome phylogenomics, biogeography, and clade diversification of Paris (Melanthiaceae). *BMC Plant Biol.* 19:543. doi: 10.1186/s12870-019-2147-6

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

Jin, W. T., Gernandt, D. S., Wehenkel, C., Xia, X. M., Wei, X. X., and Wang, X. Q. (2021). Phylogenomic and ecological analyses reveal the spatiotemporal evolution of global pines. *Proc. Natl. Acad. Sci. U.S.A.* 118:e2022302118. doi: 10.1073/pnas.2022302118

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

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

Kong, H. H., Condamine, F. L., Yang, L. H., Harris, A. J., Feng, C., Wen, F., et al. (2022). Phylogenomic and macroevolutionary evidence for an explosive radiation of a plant genus in the miocene. *Syst. Biol.* 71, 589–609. doi: 10.1093/sysbio/syab068

Kumar, V. V. N., Narayanan, M. K. R., Sunil, C. N., Sanilkumar, M. G., and Baiju, E. C. (2016). Isodon purpurescence (Lamiaceae), a new species from Western Ghats, Kerala, India. *Taiwania* 61, 13-15.

Lanfear, R., Calcott, B., Kainer, D., Mayer, C., and Stamatakis, A. (2014). Selecting optimal partitioning schemes for phylogenomic data sets. *BMC Evol. Biol.* 14:82. doi: 10.1186/1471-2148-14-82

Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., and Calcott, B. (2017). PartitionFinder 2: New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol. Biol. Evol.* 34, 772–773. doi: 10.1093/molbev/msw260

Lee-Yaw, J. A., Grassa, C. J., Joly, S., Andrew, R. L., and Rieseberg, L. H. (2019). An evaluation of alternative explanations for widespread cytonuclear discordance in annual sunflowers (Helianthus). *New Phytol.* 221, 515–526. doi: 10.1111/nph.15386

Li, H. T., Luo, Y., Gan, L., Ma, P. F., Gao, L. M., Yang, J. B., et al. (2021). Plastid phylogenomic insights into relationships of all flowering plant families. *BMC Biol.* 19:232. doi: 10.1186/s12915-021-01166-2

Li, H. T., Yi, T. S., Gao, L. M., Ma, P. F., Zhang, T., Yang, J. B., et al. (2019). Origin of angiosperms and the puzzle of the Jurassic gap. *Nature Plants.*5, 461–470. doi: 10.1038/s41477-019-0421-0

Li, H. W. (1988). Taxonomic review of Isodon (Labiatae). *J. Arnold Arbor.* 69, 289–400.

Li, H. W., and Hedge, I. (1994). "Lamiaceae," in *Flora of China*, Vol. 17, eds C. Y. Wu and P. H. Raven (St. Louis, MO: Missouri Botanical Garden Press), 196–224.

Liu, B. B., Ma, Z. Y., Ren, C., Hodel, R. G. J., Sun, M., Liu, X. Q., et al. (2021). Capturing single-copy nuclear genes, organellar genomes, and nuclear ribosomal DNA from deep genome skimming data for plant phylogenetics: A case study in Vitaceae. *J. Syst. Evol.* 59, 1124–1138. doi: 10.1111/jse.12806

Liu, M., Wang, W. G., Sun, H. D., and Pu, J. X. (2017). Diterpenoids from Isodon species: An update. *Nat. Prod. Rep.* 34, 1090–1140. doi: 10.1039/c7np00027h

Maki, M., Yamashiro, T., Dohzono, I., and Suzuki, K. (2010). Molecular phylogeny of Isodon (Lamiaceae) in Japan using chloroplast DNA sequences: Recent rapid radiations or ancient introgressive hybridization? *Plant Spec. Biol.* 25, 240–248.doi: 10.1111/j.1442-1984.2010.00290.x

Miller, M. A., Pfeiffer, W., and Schwartz, T. (2010). "Creating the CIPRES science gateway for inference of large phylogenetic trees," in *Proceedings of the Gateway Computing Environments Workshop (GCE)*, (New Orleans, LA), 1–8. doi: 10.1109/GCE.2010.5676129

Moon, H. K., Hong, S. P., Smets, E., and Huysmans, S. (2009). Micromorphology and character evolution of nutlets in tribe Mentheae (Nepetoideae. Lamiaceae). *Syst. Bot.* 34, 760–776. doi: 10.1600/036364409790139592

Murata, G., and Yamazaki, T. (1993). "Isodon (Benth.) Schrad. ex Spach," in *Flora of Japan Iiia*, eds K. Iwatsuki, T. Yamazaki, D. E. Boufford, and H. Ohba (Tokyo: Kodansha), 309–314.

Nevado, B., Atchison, G. W., Hughes, C. E., and Filatov, D. A. (2016). Widespread adaptive evolution during repeated evolutionary radiations in New World lupins. *Nat. Commun.* 7:12384. doi: 10.1038/ncomms12384

Ogishima, M., Horie, S., Kimura, T., Yomashiro, T., Dohzono, I., Kawaguchi, L., et al. (2019). Frequent chloroplast capture among Isodon (Lamiaceae) species in Japan revealed by phylogenies based on variation in chloroplast and nuclear DNA. *Plant Spec. Biol.* 34, 127–137. doi: 10.1111/1442-1984.12239

Paton, A. J., and Ryding, O. (1998). Hanceola, Siphocranion and Isodon and their position in the Ocimeae (Labiatae). *Kew Bull.* 53, 723–731. doi: 10.2307/4110492

Paton, A. J., Bramley, G., Ryding, O., Polhill, R. M., Harvey, Y. B, Iwarsson, M., et al. (2009). "Lamiaceae," in *Flora of East Tropical Africa*, eds H. J. Beentje, S. A. Ghazanfar, and R. M. Polhill (Kew: Royal Botanic Gardens), 1–431. guest.

Pham, K. K., Hipp, A. L., Manos, P. S., and Cronn, R. C. (2017). A time and a place for everything: Phylogenetic history and geography as joint predictors of oak plastome phylogeny. *Genome* 60, 720–732. doi: 10.1139/gen-2016-0191

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

R Development Core Team (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramsey, A. J., and Mandel, J. R. (2019). When one genome is not enough: Organellar heteroplasmy in plants. *Ann. Plant Rev.* 2, 1–40. doi: 10.1002/9781119312994.apr0616

Ranjan, V., Krishna, G., and Kumar, A. (2022). A new species of Isodon (Lamiaceae) from Indian Eastern Himalaya. *Taiwania* 67, 260-266.

Rautenberg, A., Hathaway, L., Oxelman, B., and Prentice, H. C. (2010). Geographic and phylogenetic patterns in Silene section Melandrium (Caryophyllaceae) as inferred from chloroplast and nuclear DNA sequences. *Mol. Phylogenet. Evol.* 57, 978-991. doi: 10.1016/j.ympev.2010.08.003

Reichelt, N., Wen, J., Pätzold, C., and Appelhans, M. S. (2021). Target enrichment improves phylogenetic resolution in the genus Zanthoxylum (Rutaceae) and indicates both incomplete lineage sorting and hybridization events. *Ann. Bot.* 128, 497–510. doi: 10.1093/aob/mcab092

Rieseberg, L. H., and Soltis, D. E. (1991). Phylogenetic consequences of cytoplasmic gene flow in plants. *Evol. Trends Plants* 5, 65–84.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., et al. (2012). MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. doi: 10.1093/sysbio/sys029

Schuster, T. M., Setaro, S. D., Tibbits, J. F. G., Batty, E. L., Fowler, R. M., McLay, T. G. B., et al. (2018). Chloroplast variation is incongruent with classification of the Australian bloodwood eucalypts (genus Corymbia, family Myrtaceae). *PLoS One* 13:e0195034. doi: 10.1371/journal.pone.0195034

Smith, S. A., Moore, M. J., Brown, J. W., and Yang, Y. (2015). Analysis of phylogenomic data sets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evol. Biol.* 15:150. doi: 10.1186/s12862-015-0423-0

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

Stegemann, S., Keuthe, M., Greiner, S., and Bock, R. (2012). Horizontal transfer of chloroplast genomes between plant species. *Proc. Natl. Acad. Sci. U.S.A.* 109, 2434–2438. doi: 10.1073/pnas.1114076109

Struck, T. H. (2014). TreSpEx-detection of misleading signal in phylogenetic reconstructions based on tree information. *Evol. Bioinform.* 10, 51–67. doi: 10.4137/EBO.S14239

Suddee, S., Paton, A. J., and Parnell, J. A. N. (2004). A taxonomic revision of tribe Ocimeae Dumort. (Lamiaceae) in continental South East Asia I. introduction, hyptidinae & hanceolinae. *Kew Bull.* 59, 337–378. doi: 10.2307/4110949

Sun, H. D., Huang, S. X., and Han, D. B. (2006). Diterpenoids from Isodon species and their biological activities. *Nat. Prod. Rep.* 23, 673–698. doi: 10.1039/b604174d

Sun, Y., Skinner, D. Z., Liang, G. H., and Hulbert, S. H. (1994). Phylogenetic analysis of sorghum and related taxa using internal transcribed spacers of nuclear ribosomal DNA. *Theor. Appl. Genet.* 89, 26–32. doi: 10.1007/BF00226978

Talavera, G., and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* 56, 564–577. doi: 10.1080/10635150701472164

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

Thiers, B. (2022). *Index herbariorum: A global directory of public herbaria and associated staff. New York botanical garden's virtual herbarium*. Available online at: http://sweetgum.nybg.org/science/ih/ (accessed March 15, 2022).

Tsitrone, A., Kirkpatrick, M., and Levin, D. A. (2003). A model for chloroplast capture. *Evolution* 57, 1776–1782. doi: 10.1111/j.0014-3820.2003.tb00585.x

Walker, J. F., Walker-Hale, N., Vargas, O. M., Larson, D. A., and Stull, G. W. (2019). Characterizing gene tree conflict in plastome-inferred phylogenies. *PeerJ* 7:e7747. doi: 10.7717/peerj.7747

Wang, J., Fu, C. N., Mo, Z. Q., Möller, M., Yang, J. B., Zhang, Z. R., et al. (2022). Testing complete plastome for species discrimination, cryptic species discovery and phylogenetic resolution in Cephalotaxus (Cephalotaxaceae). *Front. Plant. Sci.* 13:768810. doi: 10.3389/fpls.2022.768810

Wick, R. R., Schultz, M. B., Zobel, J., and Holt, K. E. (2015). Bandage: Interactive visualization of de novo genome assemblies. *Bioinformatics* 31, 3350–3352. doi: 10.1093/bioinformatics/btv383

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

Wu, C. Y., and Li, H. W. (1977). "Rabdosia (Bl.) Hassk," in *Flora reipublicae popularis sinicae*, Vol. 66, eds C. Y. Wu and H. W. Li (Beijing: Science Press), 416–534.

Xia, X. M., Yang, M. Q., Li, C. L., Huang, S. X., Jin, W. T., Shen, T. T., et al. (2022). Spatiotemporal evolution of the global species diversity of Rhododendron. *Mol. Biol. Evol.* 39:msab314. doi: 10.1093/molbev/msab314

Xiang, C. L., Dong, H. J., Landrein, S., Zhao, F., Yu, W. B., Soltis, D. E., et al. (2020). Revisiting the phylogeny of Dipsacales: New insigghts from phylogenomic analyses of complete plastomic sequences. *J. Syst. Evol.* 58, 103–117. doi: 10.1111/jse.12526

Yu, X. Q., Gao, L. M., Soltis, D. E., Soltis, P. S., Yang, J. B., Fang, L., et al. (2017). Insights into the historical assembly of East Asian subtropical evergreen broadleaved forests revealed by the temporal history of the tea family. *New Phytol.* 215, 1235–1248. doi: 10.1111/nph.14683

Yu, X. Q., Maki, M., Drew, B. T., Paton, A. J., Li, H. W., Zhao, J. L., et al. (2014). Phylogeny and historical biogeography of Isodon (Lamiaceae): Rapid radiation in south-west China and Miocene overland dispersal into Africa. *Mol. Phylogenet. Evol.* 77, 183–194. doi: 10.1016/j.ympev.2014.04.017

Zhang, R., Wang, Y. H., Jin, J. J., Stull, G. W., Bruneau, A., Cardoso, D., et al. (2020). Exploration of plastid phylogenomic conflict yields new insights into the deep relationships of Leguminosae. *Syst. Biol.* 69, 613–622. doi: 10.1093/sysbio/syaa013

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

Zhao, D. N., Ren, C. Q., and Zhang, J. Q. (2021a). Can plastome data resolve recent radiations? Rhodiola (Crassulaceae) as a case study. *Bot. J. Linn. Soc.* 197, 513–526. doi: 10.1093/botlinnean/boab035

Zhong, J. S., Li, J., Li, L., Conran, J. H., and Li, H. W. (2010). Phylogeny of Isodon (Schard. ex Benth.) Spach (Lamiaceae) and related genera inferred from nuclear ribosomal ITS, trnL-trnF region, and rps16 intron sequences and morphology. *Syst. Bot.* 35, 207–219. doi: 10.1600/036364410790862614

Zong, D., Gan, P. H., Zhou, A. P., Zhang, Y., Zou, X. L., Duan, A. N., et al. (2019). Plastome sequences help to resolve deep-level relationships of Populus in the family Salicaceae. *Front. Plant Sci.* 10:5. doi: 10.3389/fpls.2019.00005

Zwickl, D. J., and Hillis, D. M. (2002). Increased taxon sampling greatly reduces phylogenetic error. *Syst. Biol.* 51, 588–598. doi: 10.1080/10635150290102339

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 13 October 2022
DOI 10.3389/fpls.2022.990064

OPEN ACCESS

EDITED BY
Wenpan Dong,
Beijing Forestry University, China

REVIEWED BY
Jianqiang Zhang,
Shaanxi Normal University, China
Xiaoguo Xiang,
Nanchang University, China

*CORRESPONDENCE
Lihong Xiao
xiaolh@zafu.edu.cn
Jingbo Zhang
jzhang5@vcu.edu

†These authors have contributed equally to this work and share the first authorship

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

RECEIVED 09 July 2022
ACCEPTED 21 September 2022
PUBLISHED 13 October 2022

CITATION
Xi J, Lv S, Zhang W, Zhang J, Wang K, Guo H, Hu J, Yang Y, Wang J, Xia G, Fan G, Wang X and Xiao L (2022) Comparative plastomes of *Carya* species provide new insights into the plastomes evolution and maternal phylogeny of the genus. *Front. Plant Sci.* 13:990064. doi: 10.3389/fpls.2022.990064

Comparative plastomes of *Carya* species provide new insights into the plastomes evolution and maternal phylogeny of the genus

Jianwei Xi[1†], Saibin Lv[1†], Weiping Zhang[2†], Jingbo Zhang[3*†], Ketao Wang[1], Haobing Guo[4], Jie Hu[4], Yang Yang[2], Jianhua Wang[1], Guohua Xia[1], Guangyi Fan[4], Xinwang Wang[5] and Lihong Xiao[1*]

[1]State Key Laboratory of Subtropical Silviculture, Zhejiang A&F University, Hangzhou, China, [2]State Key Laboratory of Earth Surface Processes and Resource Ecology and Ministry of Education Key Laboratory for Biodiversity Science and Ecological Engineering, College of Life Sciences, Beijing Normal University, Beijing, China, [3]Department of Biological Sciences, St. John's University - Queens, NY, United States, [4]The Beijing Genomics Institute (BGI) -Qingdao, The Beijing Genomics Institute (BGI)-Shenzhen, Qingdao, China, [5]Pecan Breeding and Genetics, Southern Plains Agricultural Research Center, USDA-ARS, College Station, TX, United States

Carya, in the Juglandiodeae subfamily, is to a typical temperate-subtropical forest-tree genus for studying the phylogenetic evolution and intercontinental disjunction between eastern Asia (EA) and North America (NA). Species of the genus have high economic values worldwide for their high-quality wood and the rich healthy factors of their nuts. Although previous efforts based on multiple molecular markers or genome-wide SNPs supported the monophyly of *Carya* and its two EA and NA major subclades, the maternal phylogeny of *Carya* still need to be comprehensively evaluated. The variation of *Carya* plastome has never been thoroughly characterized. Here, we novelly present 19 newly generated plastomes of congeneric *Carya* species, including the recently rediscovered critically endangered *C. poilanei*. The overall assessment of plastomes revealed highly conservative in the general structures. Our results indicated that remarkable differences in several plastome features are highly consistent with the EA-NA disjunction and showed the relatively diverse matrilineal sources among EA *Carya* compared to NA *Carya*. The maternal phylogenies were conducted with different plastome regions and full-length plastome datasets from 30 plastomes, representing 26 species in six genera of Juglandoideae and *Myrica rubra* (as root). Six out of seven phylogenetic topologies strongly supported the previously reported relationships among genera of Juglandoideae and the two subclades of EA and NA *Carya*, but displayed significant incongruencies between species within the EA and NA subclades. The phylogenetic tree generated from full-length plastomes demonstrated the optimal topology and revealed significant geographical maternal relationships among *Carya* species, especially for EA *Carya* within overlapping distribution areas. The full-length plastome-based phylogenetic

topology also strongly supported the taxonomic status of five controversial species as separate species of *Carya*. Historical and recent introgressive hybridization and plastid captures might contribute to plastome geographic patterns and inconsistencies between topologies built from different datasets, while incomplete lineage sorting could account for the discordance between maternal topology and the previous nuclear genome data-based phylogeny. Our findings highlight full-length plastomes as an ideal tool for exploring maternal relationships among the subclades of *Carya*, and potentially in other outcrossing perennial woody plants, for resolving plastome phylogenetic relationships.

KEYWORDS

hickory, pecan, nut crop, EA-NA disjunction, structure diversity, plastome phylogeny

Introduction

The plastid, as a key uniparentally inherited organelle in plant cells, carries out not only photosynthesis but also other metabolic processes that mediate the adaptation of the plant to its environment (Dierckxsens et al., 2016). Despite increased interest in phylogenetic and phylogenomic analyses using nuclear genome datasets, plastome-based analyses have also become a powerful solution that have been widely applied to address recalcitrant relationships across the Tree of Life (Hu et al., 2016; Wei et al., 2017; Yan et al., 2018; Song et al., 2020; Wang et al., 2020; Tu et al., 2021; Dong et al., 2022; Liu et al., 2022; Ogoma et al., 2022). The plastome-based phylogenetic approaches play important roles in plant phylogenetics and evolution, such as revealing ancient and recent introgression or hybridization, tracking seed dispersal in phylogeographic studies, and charactering the structural diversity and evolution of organellar genomes (Zhao et al., 2018; Li et al., 2019; Yao et al., 2019; Li et al., 2020; Yang et al., 2021 and references therein). The most recent studies have produced numerous examples of phylogenetic discordance in plastomes or between plastids and nuclear gene trees at various evolutionary scales, which are interpreted as ancient introgressive hybridization, ancient chloroplast capture, or incomplete lineage sorting (Dong et al., 2022; Yang et al., 2021; Zhou et al., 2022). Some studies also emphasized the application of plastome-based analyses to track patterns of geographically structured interspecific gene flow, including several lineages of Fagales such as oak species with overlapping distribution ranges (Pham et al., 2017; Yang et al., 2021; Zhou et al., 2022).

Genus *Carya*, as the second largest genus in the Juglandiodeae subfamily of Fagales, contains typical temperate-subtropical forest trees for studying the phylogenetic evolution and intercontinental disjunction between eastern Asia (EA) and North America (NA) (Zhang et al., 2013; Huang et al., 2019). *Carya* is comprised of up to 20 extant species disjunctively distributed in EA and NA, and most species of the genera are also economically important for their valuable timbers and/or edible nut kernels (e.g., the pecan, Chinese hickory and Dabieshan hickory) (Grauke, 2003; Kozlowski et al., 2018; Huang et al., 2019; Xiao. et al., 2021). Based on their morphological characteristics, nineteen extant *Carya* species (excluding *C. dabieshanensis* that disputed as a variant of *C. cathayensis*) are divided into three sections based on morphological characters: section *Carya*, the true hickories (nine species); section *Apocarya*, the pecan hickories (four species); and section *Sinocarya*, the Asian hickories (six species) (Manning and Wayne, 1978; Grauke, 2003). The phylogeny and evolutionary history of *Carya* has been proposed by multiple studies based on extensive phylogenetic studies, combined with morphology, anatomy, cytology, and DNA sequences of nuclear and organelles (Manos and Stone, 2001; Manos et al., 2007; Zhang et al., 2013; Huang et al., 2019; Mu et al., 2020; Zhang et al., 2021). Zhang et al. (2013) investigated phylogenetic relationships between *Carya* species in EA and NA by integrating ten molecular markers from plastid (eight) and nuclear (two) with extensive taxon sampling, and they reconstructed the historical biogeography of *Carya* by integrating macro-fossil, morphological, and molecular data. However, the phylogeny among species in EA or NA subclades were hard to explain by any morphological traits (Zhang et al., 2013). Although the inference of biogeographical history gave more supports for the NA origin of *Carya* and the hypothetical migratory from NA to Europe and then to EA, evidence was not sufficient only based on a small set of molecular markers and partial macro-fossil records, in particular the omission of the early fossil records from the Kaliningrad region of former Soviet Union in Eurasia (Mai, 1981; Zhang et al., 2021). The most recent reports based on genome-wide SNPs (Huang et al., 2019), integrated RAD-seq and chloroplast genomes (Mu et al., 2020), and fossil-informed models provided strong support for backbone relationships

among taxa of Juglandoideae, the monophyly of *Carya* genus, and its two major subclades in EA-NA (Zhang et al., 2021). However, significant inconsistencies were found among these studies phylogenetic topologies within EA or NA subclades, especially five taxonomically disputed species, *C. poilanei*, *C. sinensis* (the synonym name is *Annamocarya sinensis*), *C. dabieshanensis*, *C. glabra* and *C. ovalis*, that inferred from different molecular markers or datasets. Furthermore, *C. illinoinensis*, as the most commercially valuable species, has shown that many cultivars of this species were generated from intraspecies hybridization with other potential admixing species, such as *C. cordiformis*, *C. aquatica* and *C. myristiciformis* (Lovell et al., 2021). Recently, based on the pan-genome assemblies of pecan genotypes and resequencing data of multiple genotypes above potential admixing species, several disease-related interspecific genomic introgression blocks have been identified in the genotypes of *C. illinoinensis* (Lovell et al., 2021). However, the maternal relationships of these pecan varieties with different economically important traits have never been addressed. Meanwhile, the maternal phylogeny of *Carya* still need to be comprehensively evaluated, and the variation of *Carya* plastome has never been thoroughly characterized.

Carya poilanei (Chev.), also known as Poilane's hickory, had three original collections in Vietnam, Laos, and Thailand (Chevalier, 1941; Leroy, 1950; Manning, 1963; Grauke et al., 1991). This rare species was initially described as *Juglans poilanei* (Chevalier, 1941). It was suspected to be extinct in the wild since the 1950s when Leroy placed the species under *Carya* (Leroy, 1950; Leroy, 1955; Grauke et al., 1991; Grauke et al., 2016). Until most recently, three subpopulations were rediscovered in the Ailao Mountain, Yunnan province, China, and *C. poilanei* was instead suggested to be listed as critically endangered (Zhang et al., 2022). Thus, the phylogenetic relationship of *C. poilanei* with other *Carya* species needs to be clarified. *Carya sinensis* (Dode), with a common name of 'Hui He Tao' (i.e. beaked walnut or beaked hickory) in China and 'Cay Cho Dai' in Vietnam, is narrowly distributed in southern China and northern Vietnam (Chevalier, 1941; Manning and Hjelmqvist, 1951). The species was first described by Dode (1912) and categorized as a separate genus *Annamocarya indochinensis* (Chevalier, 1941). It was subsequently ascribed to six different genera: *Annamocarya*, *Rhamphocarya*, *Juglandicarya*, *Caryojuglans*, *Juglans*, and *Carya* (Leroy, 1950; Manning and Hjelmqvist, 1951; Scott, 1953; Leroy, 1955; Lu et al., 1999). Although the previous and recent reports based on the molecular markers evidenced the taxon of the tree as a member of section *Sinocarya* in EA *Carya*, discordance remains among the phylogenetic topologies that were built based on different data sets (Zhang et al., 2013; Luo et al., 2021). *Carya dabieshanensis* M. C. Liu & Z. J. Li., historically considered a variant of *C. cathayensis*, is now treated as an addition to the nomenclature of Chinese hickory (Liu and Li, 1984). However, collections of germplasm and voucher specimens for more thorough comparison are necessary, and the relationship between *C. dabieshanensis* and *C. cathayensis* still needs to be addressed (Manos and Stone, 2001). Distribution maps of the widely distributed NA species, *Carya glabra* (Mill.) Sweet. with the common name pignut hickory, have included *C. ovalis* (Wangenh.) Sarg. (common name red hickory) since the reduction of *C. ovalis* to synonymy with *C. glabra* (Little, 1969). Although *C. glabra* readily hybridized with *C. ovalis* when the two occurred together, with hybrids confusing the distinctions between the species (Manning, 1950), the morphological characteristics and their habitats show that they are separate species in the section *Carya* in NA (Manning, 1950; Grauke, 2003). Nevertheless, *C. glabra* and some other species in NA *Carya* were also wrongly identified as other species partially because of the interspecific hybridization in nature stands (Heiges, 1896). These circumstances, together with the inconsistent results based on molecular inferences (Zhang et al., 2013; Huang et al., 2019), make them imperative for a more comprehensive study on the phylogeny of *Carya* and the development of molecular markers for species identification. *Carya illinoinensis* (Wangenheim) K. Koch (common name pecan), as a member of the section *Apocarya*, is native to the United States and Mexico, with over 8,000 years of history (Hester, 1983; Grauke et al., 2011; Wang et al., 2020). The latest microsatellite profiles revealed '87MX3-2.11' to be homozygous (Wang et al., 2020). Owing to the richness in health factors of its nut kernels, *C. illinoinensis* became the most commercially valuable species of *Carya*. Commercial production of *C. illinoinensis* has persisted for about a century and a half, and it has been widely planted across six continents with more than 400 cultivars released so far, including some varieties promoted and planted in large areas (Grauke et al., 2016; Xiao. et al., 2021). Among the cultivars of *C. illinoinensis*, many were formed by both natural and artificial hybridization (https://cgru.usda.gov/CARYA/PECANS). For example, the varieties 'Pawnee', 'Lakota' and 'Elliott' mentioned above, belong to intraspecies hybrids with interspecies introgression in history, based on whole genome sequencing data (Thompson et al., 2008; Huang et al., 2019; Lovell et al., 2021; Xiao et al., 2021; https://cgru.usda.gov/CARYA/PECANS). However, the maternal relationships of the varieties are still unclear, the investigation of plastome variations inherited from the female parent may provide important clues for tracking both ancient and recent gene flows among different varieties caused by inter- and intraspecies hybridization.

In this study, we newly assembled and characterized the structures and diversity of complete plastomes for the critically endangered *C. poilanei* and other 18 congeneric species, based on the sequences, generated by high-throughput sequencing technology. To explore the optimal phylogenetic relationship within *Carya*, the phylogenetic trees were reconstructed respectively by employing the entire or partial plastome sequences and unique protein-coding sequences of 30

plastomes representing all 20 *Carya* species (19 newly assembled, plus the recent release of *C. pallida*), plus six representatives from five other genera of Juglandoideae. The maternal relationship between *C. dabieshanensis* and *C. cathayensis* was determined by integrating the phylogenetic relationship with the comparison of morphological features, and confirmed the taxonomic status of *C. dabieshanensis*, *C. cathayensis*, *C. sinensis*, *C. glabra*, and *C.,poilanei* as separate species in section *Sinocarya*. The maternal relationships and discordance of species in EA and NA *Carya*, and several varieties in *C. illinoinensis* were also discussed based on the reconstruction of plastome phylogenetic topologies of *Carya*. The results generated here are of great value for the evolution, wild resource conservation and genetic breeding of *Carya* in the future. Our methodology of exploring the phylogenetic relationships of *Carya* species in EA or NA, by using the different molecular markers or the informative fragments of plastomes with different evolutionary rates, will provide a good example for the reconstruction of biogeography under subgenus in future, especially for outcrossing perennial tree species.

Materials and methods

Plant materials

In this study, fresh, fully expanded young leaves from an adult tree of *C. poilanei* (Supplementary Figure 1) were collected in July from the eastern edge of the Ailao Mountains in Jianshui County, in the southern Yunnan Province of China. Fresh, fully expanded young leaves from *C. illinoinensis* (cv. 'Sioux'), *C. cathayensis*, *C. dabieshanensis C. hunanensis*, and *C. tonkinensis* were collected from an orchard in Zhejiang A&F University, China. The samples of fresh fully expanded young leaves from *C. kweichowensis* and *C. sinensis* were collected from the Qiannan Buyi and Miao Autonomous Prefecture, Guizhou Province, China, respectively. All collected leaf samples were immediately immersed in liquid nitrogen for transportation back to the laboratory and stored at -80°C before DNA extraction. The details of the collected samples are listed in Supplementary Table 1.

The trunk, branches, leaves, flowers and fruits of *C. dabieshanensis* and *C. cathayensis* planted in the orchard of Zhejiang A&F University, Hangzhou, China, were photographed and characterized by their morphologic traits. Meanwhile, the nuts of other EA *Carya* species were also photographed.

DNA extraction and sequencing

The high-quality genomic DNA was isolated from approximately 100 mg samples using the E.Z.N.A.® HP Plant DNA Mini kit (Omega Bio-Tek, USA). DNA quality and concentration were assessed in a Qubit 3.0 Fluorometer (Thermo Fisher Scientific Inc., USA), and DNA integrity was evaluated by a 1.0% (W/V) agarose gel.

Approximately 1μg of high-quality genomic DNA from each sample was used for the whole genome sequencing (WGS). The high-throughput sequencing library of *C. poilanei* was constructed and sequenced using the Illumina NovaSeq 6000 platform (NovoGene, Beijing, China) following the standard procedure of the manufacturer. The MGIEasy DNA Rapid Library Prep Kit (cat.# 1000006985; MGI-Tech., China) was used to construct the sequencing library of *C. illinoinensis* (cv. 'Sioux'), *C. cathayensis*, *C. dabieshanensis*, *C. hunanensis*, *C. tonkinensis*, *C. kweichowensis* and *C. sinensis*, and the paired-end 100 bp reads were generated on the BGISEQ-500 platform (BGI, Shenzhen, China) according to the manufacturers' procedures, respectively. The raw sequence data was submitted to the Sequence Read Archive (SRA) of NCBI (Supplementary Table 1).

Data processes, assembly, validation and annotation

To obtain the high-quality plastome assemblies of 19 *Carya* species, 2 Gb raw reads were randomly selected for each species from NovaSeq 6000 data (*C. poilanei*), BGISEQ-500 PE100 data (seven species that were sequenced in this study) or from Illumina HiSeq X-TEN data retrieved from the NCBI database (eleven previously sequenced species), respectively. The low-quality reads from all samples, i.e., reads having over 50% bases with a quality value below 12, adaptor only and more than 10% N bases, were filtered out by SOAPnuke software (Chen et al., 2018), separately. The trimmed clean reads including nuclear and organelle genome data were used to assemble the plastomes for each species. The plastomes of 19 *Carya* species were assembled *via* NOVOPlasty software (version 3.7), with 'Seed_RUBP_cp.fasta' provided by the software as the seed input and the published *Juglans regia* plastome (accession no. NC_028617.1) as reference (Dierckxsens et al., 2016; Peng et al., 2017). The default K-mer size (K-mer = 39) was firstly selected and multiple K-mers were used for a synchronous assembling test during the assembling to obtain the ideal assemblies.

To verify the correction of the plastome assemblies, PCR amplification and Sanger sequencing were performed in the six EA species and the pecan variety 'Sioux', to further confirm the SC-IR boundaries of the assembled sequences, as well as several special genomic regions among EA species. The sequences of the primers are listed in Supplementary Table 2.

The plastome assemblies were annotated using the online program GeSeq – Annotation of Organellar Genomes (Michael et al., 2017). Initial annotation, putative starts, stops, and intron positions were predicted against the annotation of *Juglans regia*

and *C. illinoinensis* (GenBank accession numbers: NC_028617.1 and NC_041449.1) plastomes. The final annotations were determined by integrating the GeSeq prediction and manual correction. The circular plastome maps of *Carya* species were constructed using the online visualization program Organellar Genome DRAW version 1.3.1 (OGDRAW) (Stephan et al., 2019). The newly generated complete plastome sequences were deposited in GeneBank (Accession numbers were listed in Supplementary Table 3).

Comparative analyses of plastome structure features

The online program IRscope (Amiryousefi et al., 2018) was used to visualize the boundaries of LSC-IRb, IRb-SSC, SSC-IRa, and IRa-LSC of plastome among the 19 *Carya* species, The IR expansion and contraction among these *Carya* species were also compared afterward. DnaSP v. 6.12.03 (Julio et al., 2017) was employed to analyze the nucleotide diversities, sequence polymorphisms, and relative rates of plastome sequence divergence in the eighteen *Carya* species. In order to calculate the synonymous (*Ks*) and non-synonymous (*Ka*) substitution rates and the nucleotide variance (π and θ), we extracted the same individual functional protein-coding exons and aligned them separately using BioEdit v. 7.0.9.0 (Hall, 1999). To obtain selection patterns in protein-coding genes, we calculated the *Ka* and *Ks* values of each protein coding gene between two plastomes in 19 *Carya* species using DnaSP v. 6.12.03, and divided them to take the average to get the *Ka/Ks* ratio of each gene.

The online tool MIcroSAtellite (MISA, http://pgrc.ipk-gatersleben.de/misa/) was used to discover simple sequence repeats (SSRs or microsatellites) in the eighteen plastomes with the following parameters: ten for mono-nucleotide motifs, six for di-nucleotide, five for tri-nucleotide, and three for tetra-, penta- and hexa-nucleotides, respectively (Sebastian et al., 2017). The repeat numbers in the regions of LSC, SSC, IRa, and IRb were counted.

Codon usage bias was calculated using MEGA X v. 10.1.8 (Kumar et al., 2018) and EMBOSS v.6.3.1 (Peter et al., 2000; Itaya et al., 2013). In general, many genes less than 300 bp in length are believed have no real functions and the encoded proteins may not be accurate and do not necessarily contain domains. Too many such genes may interfere with the accuracy of the results. Therefore, the protein-coding genes with more than 300 nucleotides in length were extracted according to the annotation file of each plastome.

RNA-editing sites of protein-coding genes in *Carya* plastomes were predicted using the online program Predictive RNA Editor for Plant cp genes (PREP-Cp) suite with a minimal editing score of 0.7 (Mower, 2009) http://prep.unl.edu/).

Phylogenetic analyses

To construct the phylogenomic trees, 30 plastome sequences were included, including 23 plastomes from 20 species in genus *Carya* (19 newly assembled *Carya* species, 2 C. *pallida* varieties and 2 published *C. illinoinensis* varieties), 6 representative species from 5 other genera in Juglandoideae (*Juglans regia*, *J. sigillata*, *Cyclocarya paliurus*, *Engelhardia roxburghiana*, *Platycarya strobiacea*, and *Pterocarya stenoptera*), and the *Myrica rubra* in Myricaceae (as outgroups). In addition, the plastome sequences of seven other plastome assemblies were downloaded from GenBank (Supplementary Table 3). Supplementary Table 3 lists all taxa used in the phylogenomic analyses, including their sampling locations and GenBank accession numbers.

The phylogenetic relationships were inferred based on seven datasets from all 30 plastomes: the full-length sequences of plastomes; the sequences of LSC, SSC, or IR regions; the 79 concatenated protein-coding sequences (CDSs); and the LSC-SSC or LSC-IR-SSC ZZconcatenated sequences. Multiple sequence matrices of the dataset were generated in MAFFT v7.490 under standard parameters (Kazutaka and Standley, 2013), and manually adjusted, respectively. The phylogenetic trees were reconstructed based on Bayesian inference (BI) using MrBayes v3.2.6 (Fredrik et al., 2012), and the maximum-likelihood (ML) method using PHYLIP version 3.698 (http://evolution.genetics.washington.edu/phylip.html). For BI inference, two independent Markov chain Monte Carlo (MCMC) simulations were run for 2,000,000 generations and sampled every 100 generations. MrBayes settings for the best-fit model (GTR+I+G) were selected by AIC in MrModeltest. The first 25% of calculated trees were discarded as burn-in. A consensus tree and Bayesian posterior probabilities (PPBI) were constructed using the remaining trees. The ML trees were reconstructed using in PHYLIP v.7.490 and the bootstrap value was set to 500. Both BI and ML trees were visualized in FIGTREE v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/).

To facilitate our analyses, we generated a map of modern distribution for each *Carya* species. Meanwhile, we also generated a map by combining the distributions and the diversity of IR-SC boundaries with the sampling location of each analyzed *Carya* species using the ArcGIS v10.2 program (Esri. ArcGIS software on https://www.esri.com/).

Results

General characteristics of *Carya* plastomes

Using 2-Gb whole-genome resequencing data from each species, we *de-novo* assembled the complete plastomes of 19

Carya species with the *J. regia* chloroplast genome as reference (Peng et al., 2017). Boundaries between the regions of inverted repeats (IRs) and single copies (SCs) of the assemblies were verified by PCR (Supplementary Figures 2, 3; Supplementary Table 2). Sequence alignment of the amplificated sequences with our assemblies displayed a perfect match, indicating our assembly's accuracy and the inaccuracy of the assembly (additional duplication in SSC region with ca. 15 kb in length) of the published *C. kweichowensis* plastome sequence (Ye et al., 2018). Like those previously published plastomes in *Juglans* and *Carya* (Dong et al., 2017; Hu et al., 2017; Peng et al., 2017; Feng et al., 2018; Ye et al., 2018; Zhai et al., 2019; Wang et al., 2020), all the newly assembled *Carya* plastomes were circular molecules ca. 160 kb in length on average with the typical quadripartite structure, i.e., a pair of inverted repeat (IR) regions, separated by a large single-copy (LSC) region and a small single-copy (SSC) region (Figure 1; Supplementary Figure 2). Among the 19 plastomes of *Carya*, *C. poilanei* has the shortest sequence length (158,036 bp), followed by *C. tonkinensis* (158,076 bp), while *C. cathayensis* (160,823 bp) has the longest genome (Table 1). *C. cathayensis* has the longest LSC region (90,114 bp) but *C. hunanensis* has both the shortest LSC (89,468 bp) and SSC regions (18,730 bp). *C. palmeri* has the largest IR region (26,004 bp), whilst *C. poilanei* has the smallest IR region (23,634 bp) and the largest SSC region (20,861 bp). GC content do not display significant variation among the 19 *Carya* species (36.13% ~ 36.28%), as well as in the regions of LSC (33.71% ~ 33.87%), SSC (29.66% ~ 32.61%) and IRs (42.57% ~ 42.89%) among the 19 *Carya* plastomes assemblies (Table 1). However, the paired-IR regions have the highest GC content, because of the presence of four pairs of duplicated *rRNAs* in the IR regions of each species (Table 1).

The 19 *Carya* complete plastomes have the same amounts of unique genes (a total of 113 unique genes including 79 conserved protein-coding genes, 30 *tRNAs*, and 4 *rRNAs*) and introns, which were arranged with a similar gene order (Figure 1; Table 1). Fourteen of the unique genes were intron-containing genes (9 protein-coding genes and 5 *tRNA* genes), with 1 or 2 introns (Table 2; Supplementary Table 4). Five intron-containing genes (3 protein-coding genes and 4 *tRNA* genes) are located within the IR regions (Supplementary Table 4). However, the intron size varied among the *Carya* plastomes, except for *ndhB*, *rps12*, *trnA-UGC*, *trnL-UAA* and *trnV-UAC*. The length of introns in the plastomes ranged from 526 bp to 1,229 bp and the longest intron was observed in *ndhA*. Two protein-coding genes, *rpl2* and *ycf3*, exhibit significant variations in the length of both exons and introns, with considerable differences among species from EA and NA, so as the *rpoC1* and *trnA-UGC* in their intron regions.

Structural variations among *Carya* plastomes

Annotation-based circular maps of the 19 *Carya* plastomes can be categorized into three distinct structures (Figure 1). The major structure (Structure I) is shared by 16 *Carya* plastomes, with the same number of genes (131 genes, representing 113 unique genes), gene order, and translation direction (Figure 1). Eighteen unique genes were duplicated in the IR regions of these 16 plastomes (Table 2). Structure II is the most different from the major structure (Structure I) with respect to the direction of transcription of *rpl36*, which uses the complement strand as a transcriptional template in *C. sinensis* (Figure 1). A sequence loss of ~2.3 kb in length in the IRb region of *C. poilanei* and *C. tonkinensis* plastomes leads to distinguished genome structure (Structure III) and decreased gene content (a total of 129 genes). The loss of 2.3 kb sequences results in a contracted IR region and

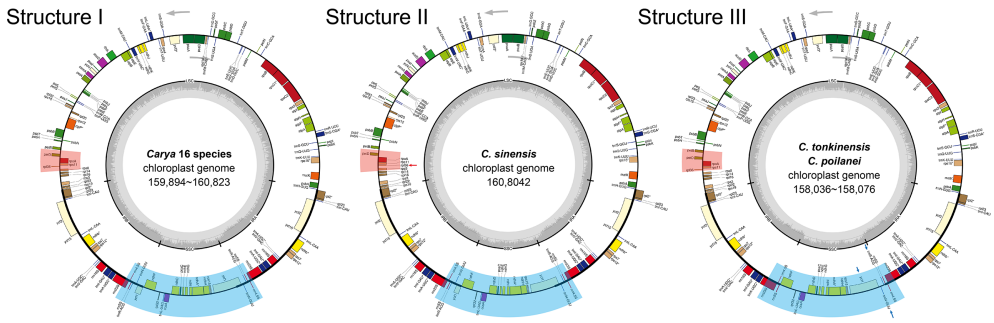


FIGURE 1
Circular maps of the 19 *Carya* species plastomes. GC content graphs are included as dark gray bars toward the center of each diagram. Intron-containing genes are marked with (*). Gray arrows indicate the translation direction of protein-coding genes. I - III show the different structures of *Carya* plastomes. Areas with light red and light blue backgrounds show differences in structure, and detailed differences are indicated by red and blue arrows.

TABLE 1 Summary of complete plastome features of the 19 newly assembled *Carya* plastomes.

Species	Length (bp)				GC content (%)				Unique genes					Total genes[d]
	Total	LSC	SSC	IRa/b	Total	LSC	SSC	IRa/b	Total	Protein coding	Intron containing[a]	tRNA[b]	rRNA[c]	
C. cathayensis	160,823	90,114	18,761	25,974	36.13	33.71	29.74	42.63	113	79	14	30	4	131
C. dabieshanensis	160,233	89,509	19,056	25,834	36.18	33.82	29.66	42.67	113	79	14	30	4	131
C. hunanensis	159,894	89,468	18,730	25,848	36.24	33.84	30.02	42.66	113	79	14	30	4	131
C. kweichowensis	160,223	89,846	18,731	25,823	36.22	33.79	30.08	42.68	113	79	14	30	4	131
C. sinensis	160,042	89,723	18,735	25,792	36.28	33.87	30.06	42.73	113	79	14	30	4	131
C. tonkinensis	158,076	89,940	20,860	23,638	36.15	33.76	31.19	42.88	113	79	14	30	4	129
C. poilanei	158,036	89,907	20,861	23,634	36.16	33.78	32.61	42.89	113	79	14	30	4	129
C. aquatica	160,763	89,966	18,791	26,003	36.15	33.75	29.89	42.58	113	79	14	30	4	131
C. cordiformis	160,793	89,989	18,798	26,003	36.15	33.74	29.88	42.58	113	79	14	30	4	131
C. floridana	160,760	89,964	18,792	26,002	36.14	33.75	29.89	42.58	113	79	14	30	4	131
C. glabra	160,652	89,888	18,786	25,989	36.20	33.80	30.02	42.57	113	79	14	30	4	131
C. illinoinensis 'Sioux'	160,714	89,917	18,791	26,003	36.16	33.76	29.89	42.58	113	79	14	30	4	131
C. laciniosa	160,787	89,921	18,864	26,001	36.16	33.79	29.81	42.58	113	79	14	30	4	131
C. myristiciformis	160,788	89,990	18,792	26,003	36.15	33.74	29.89	42.58	113	79	14	30	4	131
C. ovalis	160,822	90,032	18,786	26,002	36.15	33.74	29.89	42.58	113	79	14	30	4	131
C. ovata	160,727	89,930	18,809	25,994	36.19	33.78	29.87	42.58	113	79	14	30	4	131
C. palmeri	160,672	89,886	18,778	26,004	36.17	33.78	29.88	42.57	113	79	14	30	4	131
C. texana	160,745	89,964	18,793	25,994	36.17	33.77	29.90	42.58	113	79	14	30	4	131
C. tomentosa	160,784	89,988	18,792	26,002	36.15	33.75	29.89	42.58	113	79	14	30	4	131

a-d, the number includes only one copy if the genes are located in IR regions.

the loss of *tRNA* genes, *trnN-GUU*, and *trnR-ACG* in the IRb region (Figure 1; Table 2).

IR junction variations among *Carya* plastomes

Despite relative conservation of IR/SC boundaries in plant plastomes, contraction and expansion of the IR-SC border regions are common in the process of plastid evolution, which is the major source of variation in angiosperm-plant plastome length (Raubeson et al., 2007; Wang et al., 2018). We examined the fluctuation of IR-SC borders together with the adjacent genes in the 19 *Carya* plastomes. Of the IR-SC junctions, IRa/LSC junction has the most conserved border within a gene spacer of *rpl2* to *trnH-GUG*, with contraction sizes of 34 - 80 bp in the IRa region and 15 - 48 bp in the LSC region in the *Carya* plastomes (Figure 2A). Complete comparisons of these junctions revealed six patterns: Patterns I to VI (Figure 2A). Among them, Patterns I to IV are present in the 7 EA species and Patterns V and VI are present in the 12 NA species (Figure 2).

In details, Pattern I includes two species, *C. cathayensis* and *C. kweichowensis*, in which the LSC/IRb border is located within the *rps19* gene with the expansion of 1 bp and 4 bp from *rps19* of LSC to the IRb region, and the IRb/SSC border is located within the *ycf1* fragment with expansion 1 bp and 12 bp from *ycf1* fragment of IRb region to the SSC region (Figure 2A). Pattern II consists of *C. dabieshanensis* and *C. hunanensis*, in which the LSC/IRb border is within the gene spacer of *rps19-rpl2* with 1 bp contraction of *rps19* in the LSC region and 34 or 76 bp contraction of *rpl2* in the IRb region. The IRb/SSC border within the gene spacer of the *ycf1*-fragment and *ndhF* that contracted 5 bp and 14 bp of the *ycf1* fragment in the IRb region and 350 and 32 bp in the SSC region, and their SSC/IRa border was similar to Pattern I (Figure 2A). Pattern III includes only *C. sinensis*, which similarly has borders of LSC/IRb and SSC/IRa with the same plastomes as in Pattern II and IRb/SSC border with the same plastomes as in Pattern I. Pattern IV contains *C. poilanei* and *C. tonkinensis* plastomes, with a distinguished IR-SC boundary pattern: IRb/SSC border within the *ndhF* gene with a 6-bp expansion from the SSC region to the IRb region, and a SSC/IRa border located in the gene spacer between *trnR-ACG* and *rrn5S* with contractions of 178 bp and 77 bp for both sides, respectively (Figure 2A). Meanwhile, the LSC/IRb border of the plastomes in Pattern IV is similar to those in Pattern II (Figure 2A). Of the 12 NA *Carya* species, 11 have the highly conserved SC-IR boundaries and were assigned into Pattern V, which has the LSC/IRb border within *rps19* with 2 bp or 4 bp expansion to the IRb region, the IRb/SSC border on the last base of

10.3389/fpls.2022.990064

TABLE 2 Genes and contents in the newly assembled *Carya* plastomes.

Category	Group	Genes and number		
		LSC	SSC	IRa/b[a]
Photosynthesis	Subunits of photosystem I	*psaA, psaB, psaI, psaJ* (4)	*psaC* (1)	–
	Subunits of photosystem II	*psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ* (15)	–	–
	Subunits of NADH dehydrogenase	*ndhC, ndhJ, ndhK* (3)	*ndhA*, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI* (7)	*ndhB** (1X2)
	Subunits of cytochrome b/f complex	*petA, petB, petD, petG, petL, petN* (6)	–	–
	Subunits of ATP synthase	*atpA, atpB, atpE, atpF*, atpH, atpI* (6)	–	–
	Large subunit of Rubisco	*rbcL* (1)	–	–
Self-replication	Large subunits of ribosome	*rpl14, rpl16, rpl20, rpl22, rpl33, rpl36* (6)	*rpl32* (1)	*rps7, rps12**[d] (2X2)
	Small subunits of ribosome	*rps2, rps3, rps4, rps8, rps11, rps14, rps16*, rps18, rps19* (9)	*rps15* (1)	*rpl2*, rpl23* (2X2)
	DNA-dependent RNA polymerase	*rpoA, rpoB, rpoC1*, rpoC2* (4)	–	–
	Ribosomal RNAs	–	–	*rrn16S, rrn23S, rrn4.5S, rrn5S* (4X2)
	Transfer RNAs	*trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnfM-CAU, trnG-GCC, trnH-GUG, trnK-UUU, trnL-UAA*, trnM-CAU, trnP-UGG, trnQ-UUG, trnR-UCU, trnS-CGA*, trnS-GCU, trnS-GGA, trnS-UGA, trnT-GGU, trnT-UGU, trnV-UAC*, trnW-CCA, trnY-GUA* (22)	*trnL-UAG* (1)	*trnA-UGC*, trnI-CAU, trnI-GAU*, trnL-CAA, trnN-GUU*[b], *trnR-ACG*[b], *trnV-GAC* (7X2)
Other genes	Maturase	*matK* (1)	–	–
	Protease	*clpP** (1)	–	–
	Envelope membrane protein	*cemA* (1)	–	–
	Acetyl-CoA carboxylase	*accD* (1)	–	–
	C-type cytochrome synthesis gene	–	*ccsA* (1)	–
Hypothetical reading frames	Proteins of unknown function	*ycf3*, ycf4* (2)	*ycf1*[c] (1)	*ycf15, ycf2* (2X2)

a, located in IR regions with two copies; b, *trnR-ACG* and *trnN-GUU* only located in SSC region in *C. tonkinensis* and *C. poilanei* with single copy; c, *ycf1* gene located in SSC region for *C. tonkinensis* and *C. poilanei*, but the most part of its genic region located in SSC and small part in IRa region for other species; d, the most part of rps12 located in IR regions and only small part in LSC region; * indicated intron-containing gene.

the *ycf1*-fragment in IRb and SSC/IRa border within the *ycf1* gene with 1,093 bp expansion to IRa region (Figure 2A). Pattern VI includes only *C. ovata*, with the same borders of LSC/IR and SSC/IRa, except for the IRb/SSC border, which was within the *ycf1*-fragment with 5-bp expansion to the IRa region (Figure 2A). These results indicated more conserved SC-IR boundary patterns among NA *Carya* plastomes than those in EA that exhibit highly diverse IR-SC boundary patterns (Figure 2B).

10.3389/fpls.2022.990064

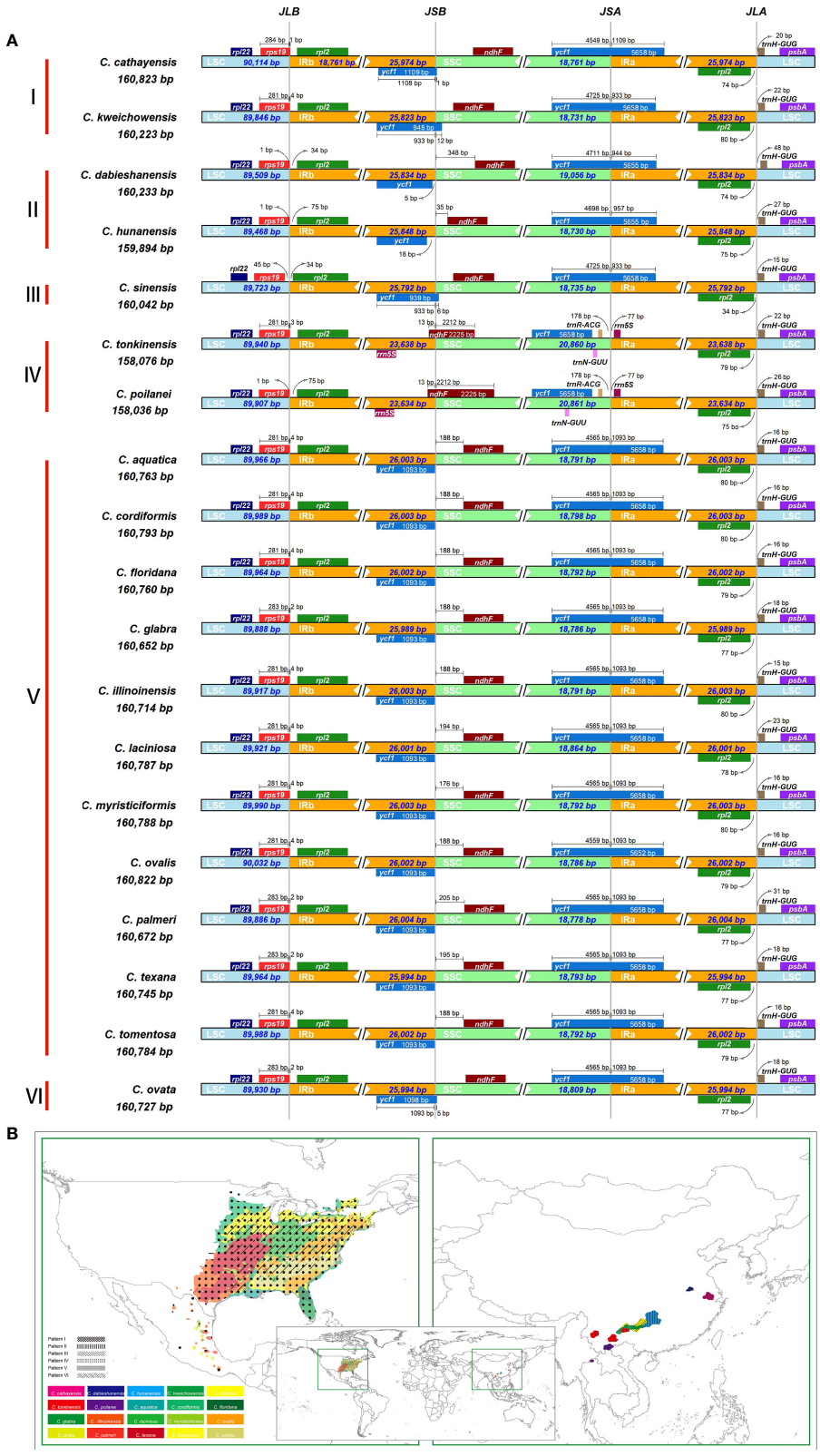


FIGURE 2
IR junction variation and its diversity among the 19 newly assembled *Carya* plastomes. **(A)**, Comparison of IRb-LSC, LSC-IRa, IRa-SSC, and SSC-IRb boundaries. **(B)**, The modern distribution and the diversity of IR-SC boundaries.

10.3389/fpls.2022.990064

Simple-sequence repeats among *Carya* plastomes

Simple-sequence repeats (SSRs), also known as microsatellites, are short tandem repeat DNA sequences that consist of repeating 1-6 nucleotide motifs widely distributed throughout the plastomes, which are important molecular markers for analysis of genetic diversity and relationships between species (Yang et al., 2011; Jiao et al., 2012). We detected 1,652 SSRs in the 19 *Carya* plastomes and the numbers of SSRs varied among the species, with a range from 76 in *C. aquatica* and *C. texana* to 106 in *C. poilanei* and *C. tonkinensis* (Figure 3A; Supplementary Table 5). Of these detected SSRs, mononucleotide repeats were the most abundant SSR motifs, which accounted for approximately 85.5% of the total SSRs, and over 99% of the mononucleotide repeats were composed of A/T repeats (Figure 3A; Supplementary Figure 4). While the tri- to hexanucleotide SSRs were occupied 7.75% of the total (Figure 3A; Supplementary Figure 4). In addition, the total number of SSRs in each *Carya* plastome also revealed significantly different patterns among species in EA (93 - 106) and NA (76 - 82) *Carya* species, in which the mononucleotide repeats (especially for A/T repeats) are the

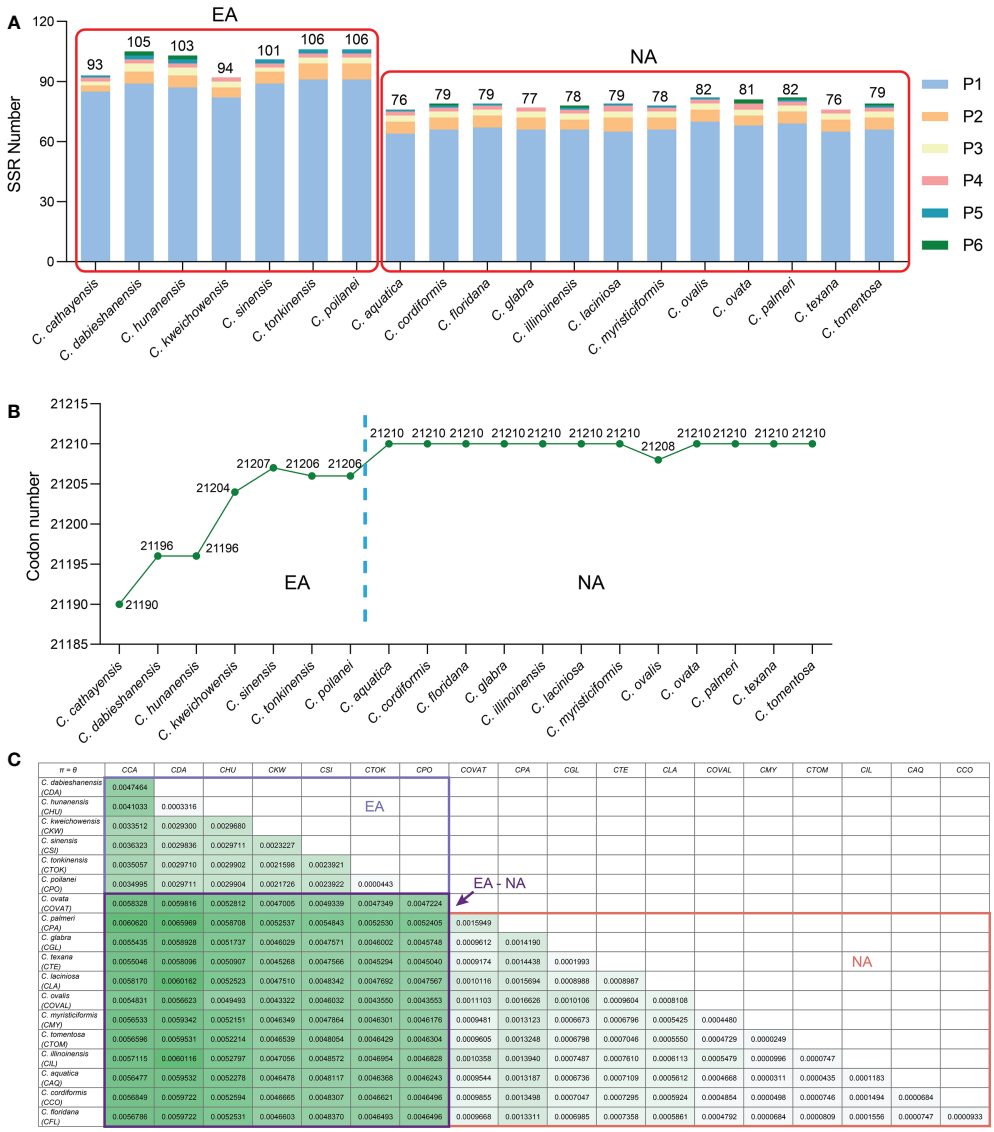


FIGURE 3
Comparison on SSR numbers, codon number and nucleotide diversities of *Carya* plastomes. **(A)**, Statistics of SSR number for each *Carya* plastome. **(B)**, Statistics of codon number for each *Carya* plastome. **(C)**, The paired nucleotide diversities of *Carya* plastomes from species in Eastern Asia (EA) or North America (NA), or between two species from Eastern Asia and North America (EA-NA).

major contributor (Figure 3A; Supplementary Table 5). Meanwhile, the numbers of SSRs of the plastomes can also be classified into six patterns, corresponding to those of IR-SC boundary patterns (Figures 2, 3A).

Codon usage analysis

Codon distribution analysis of 52 protein-coding genes with a length over 300 bp for each gene sequence in the *Carya* plastomes revealed a similar codon usage pattern among the plastomes of *Carya* species and *J. regia* (Supplementary Table 6). The protein-coding sequences contain a total of 21,190 to 21,210 codons in the *Carya* plastomes and 21,265 in *J. regia*, including stop codons. The total number of codons varied among EA *Carya*, whereas 11 of the NA *Carya* have the same number of total codons (21,210), with the only exception of NA species in *C. ovalis* plastome with 21,208 codons (Figure 3B). The total number of codons for each species can also be classified into six patterns as mentioned above.

Other than the codon distribution, we also counted the average effective numbers of codon (ENC) in the plastomes. The results showed that 17 of the 19 assembled *Carya* plastomes had similar codon usage bias (ENC ranged from 48.883 to 49.001), which were slightly lower than that of *J. regia* (48.776) (Supplementary Table 6). *C. tonkinensis* and *C. poilanei* have ENC = 56.51 and 56.52, which are significantly higher than any of other species. All the *Carya* plastomes have similar GC content in the codons (37.32% ~ 37.39%) except for *C. tonkinensis* and *C. poilanei* plastomes which have lower GC contents for the first (39.59%) and second codons (31.12% and 31.18%) and relatively high GC content (41.26% and 41.20%) for the third position (Supplementary Table 6). Among all codons in the 19 *Carya* plastomes, leucine is the most abundant amino acid, while cysteine showed the least abundance (Supplementary Figure 6A). The most frequent synonymous codon in the *Carya* plastomes was AUU, which encodes isoleucine, and the least frequent codon, except for the stop codons, is UGC, which encodes cysteine (Supplementary Figures 5, 6B). Of the 20 amino acids and stop codons, only leucine each has six codon types, while methionine and tryptophan preferred one codon type in the plastomes of 19 *Carya* species and *J. regia* (Supplementary Figures 6). In general, AUG encodes not only methionine but also the universal start codon for the nuclear genome of eukaryotes (Thach et al., 1966). We detected that the genes of *ndhD*, *rpl16*, *ycf15*, and *rps19* use ACG, ATC, CTG, and GTG encodes (not ATG) to encode the start codon – AUG respectively.

We also assessed the relative synonymous codon usage (RSCU), a good indicator for measuring nonuniform synonymous codon usage bias in coding sequences (Lee et al., 2010). Both methionine and tryptophan have RSCU = 1 (Supplementary Figure 7), indicating that codons AUG and UGG have no bias or preferences. The plastomes of 19 *Carya* species and *J. regia* had 30 biased codons with RSCU > 1, and their third position is A/U except for leucine (UUG). UUA has the highest RSCU values (1.94 to 2.03) in leucine in all the plastomes of *Carya* species and *J. regia*, and AGC has the lowest RSCU values in leucine (0.38 for *Carya* species and *J. regia*). Moreover, leucine showed A or T (U) bias in all synonymous codons: UUA, UUG, CUU, CUC, CUA, and CUG. We observed that the RSCU value for the specific amino acid increased with the number of codons. The following codons have high RSCU frequency (>30%) and fraction: GAU (aspartate), GAA (glutamate), AUU (isoleucine), AAA (lysine), UUA (leucine), AAU (asparagine), UAU (tyrosine), UUU (phenylalanine), and CAA (glutamine); and the bias toward these nine codons was consistent with the low content of GC in the third codon position (Supplementary Figure 7). Our analysis also showed that the RSCU value increased with the quantity of codons coding for a specific amino acid.

Prediction of RNA editing

We found that 62 post-transcriptional RNAs have editing modifications in 23 protein-coding genes (Supplementary Figure 9; Supplementary Table 7). Genes of *ndhB* (11 editing sites) and *ndhD* (10 editing sites) contain the most RNA editing sites, followed by *rpoB* and *rpoC2*, with 6 and 5 editing sites respectively. Only one to three RNA editing sites were found in the rest of the 19 protein-coding genes. All of RNA editing sites potentially caused the conversion from cytosine to uracil after transcription, and 43 of them (69.4%) took place at the second nucleotide of codons, with two times the conversion rate compared to the first nucleotide (19, 30.6%) (Supplementary Figure 8). However, no correlation was observed between gene length and gene number (Supplementary Figure 8). Approximately 77% of the RNA modifications resulted in the conversion of hydrophilic to hydrophobic amino acids, mainly serine to leucine or phenylalanine (S to L, 22 editing sites; S to F, 8 editing sites), or proline to leucine (P to L, 8 editing sites; Supplementary Table 7). We also observed the conversion from proline to serine in three editing sites, representing the changes of amino acids from nonpolar to polar. On the species level, the majority of the RNA-editing sites were shared by all the *Carya* plastomes and only a few were species-specific, for example, CUC to UUU (L to F) was found only in *C. poilanei* at the amino acid position 71, CGG to UGG (R to W) only in *C. sinensis*, CCA to UCA (P to S) only in *C. floridana*, and CUU to UUU (L to F) only in *C. laciniosa* (Supplementary Table 7).

Analyses of selection and nucleotide diversity

In order to obtain selection patterns in protein-coding genes, the nonsynonymous (*Ka*) and synonymous (*Ks*) substitutions

ratios (*Ka/Ks* = ω) were determined for 79 unique protein-coding genes with the comparison of the 19 *Carya* plastomes (Supplementary Figure 9; Supplementary Table 8). Among the protein-coding genes, *ndhA* and *petA* have *Ka/Ks* ratios greater than 1.0 (Supplementary Figure 10; Supplementary Table 8), indicating the genes were subjected to positive selection. The rest of the 77 protein-coding genes have *Ka/Ks* < 1.0 (Supplementary Figure 9; Supplementary Table 8), indicating that purification selection happened only during plastome evolution. Specially, seven genes (*atpF*, *ccsA*, *matK*, *rbcL*, *rpoA*, *rps15* and *ycf1*) have the *Ka/Ks* ratios between 0.5 and 1.0 (Supplementary Figure 9; Supplementary Table 8). These results clearly indicate that protein-coding genes in plastomes of different plant species were subjected to diverse selection pressures.

Two parameters, Pi (π) and theta (θ), were used for measuring the nucleotide variability among protein-coding genes of the 19 *Carya* plastomes. The value πvaried among the protein-coding genes with a range from 0 to 0.08495 (average π = 0.00326) (Supplementary Table 8). The gene *rpl36* (π = 0.08495) was notably variable among the protein-coding genes in the 19 *Carya* plastomes. The gene *ycf1* showed a relatively lowerπ value (0.00493) among the protein-coding genes in the *Carya* plastomes, although it was commonly used as a representative plant DNA barcoding region (Dong et al., 2015). The variation of the other parameter θ showed the same patterns as the π value among *Carya* plastomes of the 79 protein-coding genes (Supplementary Table 8). In comparison, plastomes of EA *Carya* species revealed higher nucleotide diversities (π and θ) than those in NA *Carya* species (Figure 3C).

Maternal phylogenetic inference within *Carya*

The phylogenetic analyses, based on matrices (with indels) from seven datasets containing the full or partial sequence of 30 plastome sequences including 26 Juglandaceae species and one *M. rubra* (as tree root), resulted in seven different topologies using both Bayesian inference and the ML method (Figure 4; Supplementary Figure 10–15). Among the seven BI trees, one generated from the IR-region dataset displayed a very chaotic phylogenetic relationship among all analyzed species and varieties, and therefore was discarded for further analysis (Supplementary Figure 15). The remaining six BI trees displayed identical topologies (with posterior probability (PP) value = 1) on the intergeneric level in Juglandiodeae. All six topologies supported the two sister subclades of EA and NA in the genus *Carya* (Figure 4; Supplementary Figures 10–14).

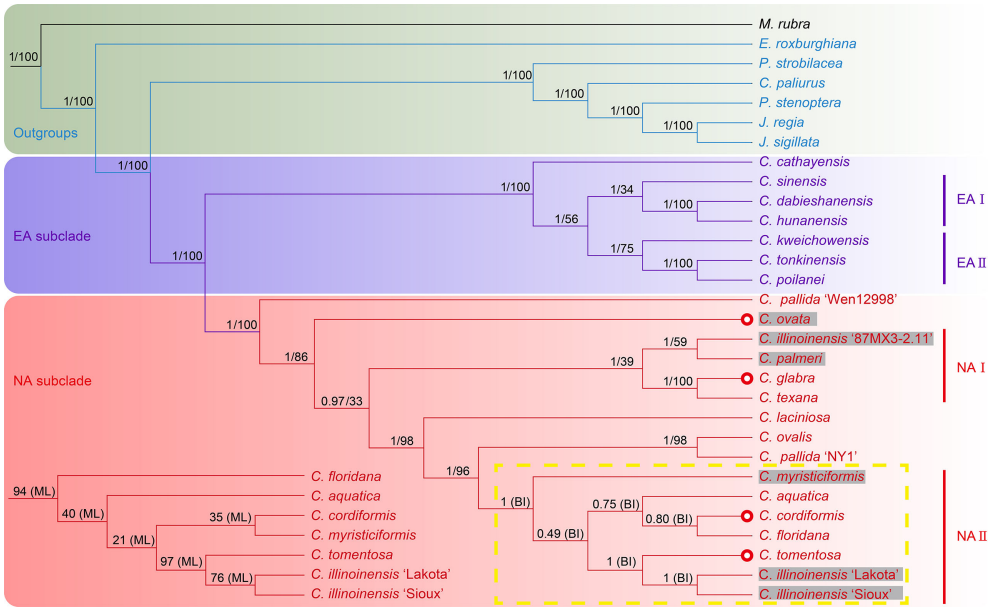


FIGURE 4
Phylogenetic trees of *Carya* species and representative species in other genera of Juglandaceae inferred from the datasets of complete plastomes using Bayesian inference and maximum likelihood method. The posterior probability (PP) and bootstrap (BS) values that supported each node are shown above the branches. The colors in green, light purple and pink indicate the outgroups, EA and NA *Carya* species, respectively. The letter n represents the haplotype, and x represent and the number before it shows the chromosomal set of the haplotype of each species. The numbers after n and x are the chromosomes of haplotype of each species. Red circles show the species in section *Apocarya*, gray background and the red circles indicate the tropical distributed species/varieties. The lower left quarter and the yellow box respectively show the parts with different topologies in the trees based on ML and BI methods.

When considering the overall supporting rates for the six topologies based on BI, we are confident that the best phylogenetic tree is generated by full-length plastome sequence, followed by LSC, LSC-SSC, SSC, CDS, and LSC-IR-SSC (Figure 4; Supplementary Figures 10–14). Almost all branches of the BI tree inferred from full-length plastomes displayed the highest supports of any of the other 5 phylogenetic trees based on Bayesian inference. The topologies of all ML trees highly supported the backbone relationships among genera of Juglandoideae and the two sister subclades of EA and NA in *Carya* (Figure 4; Supplementary Figures 10–15). The overall supporting rates were in the same sequence as mentioned above for BI trees (Figure 4; Supplementary Figures 10–15). Excluding the trees based on the IR dataset, the genus *Engelhardia* formed a sister relationship with five other genera (including *Carya*) with full supports (PP = 1 or BS = 100) in the remaining partial or full-length plastome sequences-based topologies generated by Bayesian inference and the ML method (Figure 4; Supplementary Figures 10–14). These topologies fully supported the hierarchical clustering of *Platycarya*, *Cyclocarya*, *Pterocarya*, and *Juglans*, and grouped a sister clade with *Carya*. Within the *Carya* monophyletic group, two major subclades, i.e., EA and NA are well-separated (Figure 4; Supplementary Figures 10–14). However, the supporting rates [bootstrap (BS) values] for the branches in all seven ML trees were significantly lower than the PP values in BI trees, and the topologies in the EA and NA subclades varied among the ML trees (Figure 4; Supplementary Figures 10–15). Compared to the corresponding BI tree, similar phylogenetic topologies were obtained using the ML method based on the data set generated by full-length plastome sequences, which also represented the best ML tree although with relatively low supports (BS values) of most branches (Figure 4; Supplementary Figures 10–15). Therefore, the BI trees are considered the primary references in the following analyses of plastome phylogenetic relationships of *Carya*, and the detailed analyses were mainly based on the BI topology inferred by full-length plastome sequences (Figure 4; Supplementary Figures 10–14).

The plastome phylogenetic relationships inferred from datasets of full-length, LSC, and LSC-SSC regions of plastomes displayed the same topologies for EA subclade with PP values of 0.997 (LSC BI phylogenetic tree) to 1 (full-length and LSC-SSC BI trees) (Figure 4; Supplementary Figure 10). However, the maternal phylogenetic incongruencies were obtained between species of EA or NA *Carya* in the topologies inferred from partial or full-length plastome datasets. In contrast, the maternal phylogenetic trees inferred from the datasets of SSC, CDS, and LSC-IR-SSC showed completely different topologies for EA *Carya* subclades, with relatively low PP values for several internal subclades (Supplementary Figures 11–13). However, all six BI topologies support the sister relationships for two pair of terminal branches, *C. dabieshanensis* – *C. hunanensis* and *C. tonkinensis* – *C. poilanei*, with supports = 1 (Figure 4; Supplementary Figures 11–13). Within the EA subclade, *C. cathayensis* was considered the base taxon with PP = 1 in four topologies of full-length, LSC-SSC, SSC and LSC-IR-SSC. Meanwhile, the topologies of full-length and LSC fully supported the sister relationships with the nodes of *C. sinensis* – *C. hunanensis* – *C. dabieshanensis* (designated as EA-I) and *C. kweichowensis* – *C. tonkinensis* – *C. poilanei* (designated as EA-II) (Figure 4; Supplementary Figures 10–13). *C. kweichowensis* formed a sister node with the terminal branch of *C. tonkinensis* – *C. poilanei* in four topologies of full-length, LSC, LSC-SSC, and SSC of plastomes. *C. sinensis* formed a sister relationship with the terminal branch of *C. dabieshanensis* – *C. hunanensis* in the phylogenetic trees of full-length plastomes and LSC-SSC sequences. This relationship was strongly supported in the LSC dataset-based maternal phylogeny but formed a terminal sister relationship with *C. kweichowensis* in the SSC dataset-based phylogenetic tree. Although *C. dabieshanensis* and *C. cathayensis*, have close geographical distribution and similar morphological traits, the former has larger size of fruits and nuts than the latter, and their maternal phylogenetic relationship is relatively far away (Figure 4; Supplementary Figure 17).

In the NA subclade, five out of six BI topologies (except for that inferred from SSC dataset) fully supported the *C. pallida* var. 'Wen 12998' as the basal taxon, while *C. pallida* variety 'NY1' formed a terminal branch with *C. ovalis* (Figure 4; Supplementary Figures 10–13). All six BI topologies fully supported the terminal branch of *C. glabra* and *C. texana*. The variety of *C. illinoinensis* '87MX3-2.11' displayed a sister relationship with *C. palmeri* in the full-length and SSC-based phylogenetic trees, and formed a sister relationship with *C. laciniosa* in the LSC, LSC-SSC, CDS, and LSC-IR-SSC-based phylogenetic trees. The other two varieties of *C. illinoinensis* – 'Lakota' and 'Sioux' formed a terminal branch with very strong supports and a sister branch to *C. tomentosa* in the full-length, LSC, LSC-SSC, and LSC-IR-SSC-based maternal phylogenetic trees. *C. aquatica* and the terminal branch of *C. cordiformis* – *C. floridana* formed a sister note in full-length and LSC plastome sequences-based maternal phylogenetic trees. The phylogenetic relationship of *C. ovata* varied in five phylogenetic trees (full-length plastomes, LSC, LSC-SSC, SSC, and CDS-based inferences).

In summary, the maternal phylogenetic relationships within the subclades EA or NA were clear based on the topology inference of complete plastomes with high confidence (PP > 0.8) for most nodes and the topology was selected for further discussion on the maternal phylogenetic relationships within EA or NA subclade (Figure 4). The topology of the EA subclade fully supported the earliest divergence of *C. cathayensis*, which consists of a sister relationship with two sub-sister nodes, *C. sinensis* – *C. hunanensis* – *C. dabieshanensis* (EA-I) and *C. kweichowensis* – *C. tonkinensis* – *C. poilanei* (EA-II), indicating that they came from a common ancestor in the genus. Among the species of the NA subclade, one landrace 'Wen 12998' of *C.*

pallida was clustered as a basal taxon and formed a sister relationship with the rest of the NA plastomes. Except for the node in NA-II (PP = 0.502) containing *C. tomentosa*, two cultivars of *C. illinoinensis*, and the sister node (PP = 0.755) of *C. aquatica*, *C. cordiformis* and *C. floridana*, the majority nodes of the NA subclade topology displayed very high supporting rates. The topology of NA subclade disrupted the morphological features-based division of the two sections, *Carya* and *Apocarya* (Huang et al., 2019), but varieties that are at the same terminal branch or that have a close relationship in the phylogenetic tree indicate that they have close or overlapping distribution regions (Figures 2B, 4; Supplementary Figure 17). Two pairs of terminal branches containing a local collection '87MX3-2.11' of *C. illinoinensis* (Wang et al., 2020), and *C. palmeri* (section *Carya*) section and species *C. glabra* and *C. texana* (section *Apocarya*) consisted of an internal sister node (NA-I) with *C. ovata* (section *Carya*) in the NA subclade (Figure 4). In contrast, *C. myristiciformis*, as the partial sister of the terminal branch of *C. ovalis* and the *C. pallida* var. 'NY1', showed a relatively distant maternal phylogenetic relationship with the two species (Figure 4). Interestingly, two varieties of *C. illinoinensis* – 'Sioux' and 'Lakota' which had different parentages and came from different controlled crosses made in Brownwood, Texas in 1943 and 1964, respectively (Grauke and Thompson, 1996; Wang et al., 2020), also showed distant phylogenetic relationships with the wild seedling '87MX3-2.11' that originated from Oaxaca, Mexico (Wang et al., 2020) (Figure 4).

Discussion

Characteristics and comparison of *Carya* plastomes: Conservation and diversity

Being consistent with reports in most flowering plants, *Carya* plastomes have conservative canonical quadripartite structures including LSC and SSC regions (Figure 1), which are separated by two IR regions that are generally known to play a role in the structural stability of plastomes (Palmer and Thompson, 1982; Bock, 2007). The *Carya* plastomes are similar in gene content and gene order (113 unique genes) (Figure 1; Tables 1, 2), as reported in many angiosperms (Hu et al., 2016; Xu et al., 2017; Mader et al., 2018; Yang et al., 2018). A strong tendency toward A or U at the third codon position (Clegg et al., 1994; Gao et al., 2017; Mader et al., 2018; Meng et al., 2018) may explain why the A/T content is as high as ~ 64% in *Carya* plastomes (Table 1). The conservatism of *Carya* plastomes is also revealed by RNA editing sites and the measurement of selection for each protein-coding gene (Supplementary Table 7).

As the most conserved region in the plastomes, IR expansion or contraction may alter the plastome size and the stability of the genomic structure, which in turn could cause IR-SC boundary shift, repeated sequences (Dugas et al., 2015), and/or duplication or deletion of a certain gene through inversion during evolution (Park et al., 2018). Frequent expansions and contractions at the junctions of IR-SCs have been recognized as evolutionary signals for illustrating the relationships among taxa (Khakhlova and Bock, 2006; Raubeson et al., 2007; Wang et al., 2008; Lu et al., 2018; Park et al., 2018). Moreover, repeated sequences, especially SSRs play key roles in plastid genome rearrangement, divergence, and evolution (Weng et al., 2014). Our analyses reveal the intercontinental disjunctive distribution between and among EA and NA *Carya* species, reflected by the patterns of IR-SC junctions, the number of SSRs and codons, and the paired nucleotide diversities of *Carya* plastomes (Figures 2, 3). The diversified features described here for *Carya* plastomes and the shortened IR regions in *C. tonkinensis* and *C. poilanei* could be related to the variation in genome size and the long evolutionary history of *Carya* species. By comparison, *Carya* species in EA exhibited higher diversities in plastome features and structure compared to those in NA (Figures 1–3; Table 1), although higher diversities in species and morphology were displayed in NA *Carya* than those in EA (Huang et al., 2019). The interesting findings combined with geology and vicariance events suggest more diverse matrilineal sources of EA *Carya* species, in contrast to fewer matrilineal sources of NA *Carya* species during their evolution. Meanwhile, the changes in climates and habitats especially in the glacial period may speed up the adaptive speciation in NA *Carya* and lead to higher speciation and morphological diversities in NA than those in EA. It is known that climate cooling is commonly accepted as the main causes of the isolated habitats and the disjunctions of floristic elements between continents (Raven, 1972; Tiffney, 1985; Wen, 1999). One example as addressed by Deng et al. (2017), is the modern Asia-North America disjunct distribution of two taxa from Rubiaceae – *Kelloggia* and *Theligonum* formed by the fragmentations of original wide distribution caused by climate cooling. Extinction of *Carya* in EA and NA might occur commonly during climate change. The rapid uplift of Qinghai-Tibetan Plateau and climate change could be responsible for the modern restricted distribution of *Carya* in EA (Supplementary Figure 16), while the overlapped distribution area and frequent inter- and intra-species hybridization could account for the relatively low plastome diversity in modern NA *Carya* species (Supplementary Figure 16).

Complete plastomes: An excellent tool for inferring the matriarchal phylogeny and geography of *Carya*

Our maternal phylogenetic analyses showed that six of the BI-based and all ML-based topologies inferred by different datasets of plastomes support the sister relationship of *Carya* and the monophyletic *Cyclocarya-Juglans-Platycary-Pterocarya*

clade in Juglandoideae (Figure 4; Supplementary Figures 10–13). Our results also strongly support two monophyletic subclades corresponding to the disjunctive geographical distributions of *Carya* in EA and NA (Figure 4; Supplementary Figures 10–13). The results are highly consistent with those of previous inferences on the intergeneric phylogenetic relationships in Juglandoideae and the relationships between two subclades in the genus *Carya*, based on the molecular markers from partial or complete nuclear or/and organelle sequences (Manos and Stone, 2001; Manos et al., 2007; Zhang et al., 2013; Zhang et al., 2019; Mu et al., 2020; Zhang et al., 2021; Zhou et al., 2021). It is known that the plastome is a key organelle in the plant cells, performing photosynthesis and other metabolic processes related to the adaptation of the plant to its environments (Dierckxsens et al., 2016). Although the basic structure of the plastome is highly conserved throughout land plant lineages, it has been proven that differences in the sizes of the complete genome and the protein-coding gene content of the different genome regions related to the difference in selection pressure are informative in phylogeny and evolution for many plant lineages. Although the matriarchal phylogenetic relationships among species and/or varieties within the EA and NA subclades varied among the topologies inferred from different datasets of plastomes, the plastome-based physiologies still provide very important clues to the backbone relationships between EA and NA *Carya* (Figure 4; Supplementary Figures 10–13). The plastome-based phylogeny of *Carya* could also provide a good example for exploring the evolution of plastomes as we present in this study.

By comparing the PP and BI values of all nodes among six topologies, almost all the nodes of the BI tree reconstructed from full-length plastomes showed the highest support (PP > 0.8) (Figure 4; Supplementary Figures 10–13). Therefore, the full-length plastome dataset-based BI tree is considered the main reference for further analyses of maternal phylogenetic relationships among *Carya* species. The maternal phylogenetic relationships among EA *Carya* species are highly in agreement with their geographical distributions (Figures 2B and 4; Supplementary Figure 17). The maternal phylogenetic positions of the species in NA subclade provide strong support for their overlapping and adjacent distribution patterns (Figures 2B and 4; Supplementary Figure 17), although the topology does not match the morphological features-based division of two sections (Manos and Stone, 2001; Huang et al., 2019). Meanwhile, the topologies among species within EA and NA subclades exhibited significant deviation from the pattern presented by nuclear genome data (Huang et al., 2019). Our results demonstrated that the different contributions of nuclear and organelle genomes, compared to the control, based on the morphological features associated with the different evolutionary rates between organelle and nuclear genomes, showed the difference between topologies based on the different datasets of plastomes in the present study. In general, the nonparental-inherited plastomes are highly conservative in protein-coding gene functions and genome structure, which can also reveal the maternal origin and diversity of a plant (Raubeson and Jansen, 2005). However, the morphological characteristics could be controlled by parental-inherited materials in the nucleus and may be influenced by environment and selection resulting in adaptation, convergent evolution, and species diversity (Givnish, 2016). Taking this into account, it is easy to understand the contradiction between topologies built from plastomes and nuclear data, as both had different evolutionary trajectories.

Carya poilanei, has been regarded as an extinct species in the subfamily Juglandoideae for more than 60 years before its rediscovery in the Ailao Mountain area, Jianshui County, Yunnan Province, China (Zhang et al., 2022). This species has historically been classified as a member of Juglans (Chevalier, 1941), then moved to the genus *Carya* (Leroy, 1950), and finally botanically characterized (Leroy, 1955). Although it is classified in the section *Sinocarya* of the genus *Carya* based on its morphological traits, the phylogenetic and maternal relationships are still unclear. The rediscovery of the species brings this issue back into consideration. Our results indicate that *C. poilanei* has the highest similarity with *C. tonkinensis* in plastome structure and features, and it has the closest maternal phylogenetic relationship with *C. tonkinensis*, which is supported by their closest geographical distribution (Figures 1–4; Tables 1, 2; Supplementary 17). Therefore, we assume that these two species historically shared a common maternal ancestor. In addition, the taxonomic placement of four contentious species *C. dabieshanensis*, *C.* (*Annamocarya*) *sinensis*, *C. glabra*, and *C. ovalis* has been disputed for decades (Chang and Lu, 1979; Thompson and Grauke, 1991; Manos and Stone, 2001; Grauke and Mendoza-Herrera, 2012; Grauke et al., 2016). *C. dabieshanensis* has been treated as a member of *C. cathayensis* because of their high similarities in morphological features, except for the larger fruit size of *C. dabieshanensis* (Liu and Li, 1984). Our previous nuclear genome analyses also supported that *C. dabieshanensis* is close to *C. cathayensis*, but not to any members in the section Sinocarya (Huang et al., 2019). This study provided a full comparison of the morphological features between these two species and found no additional differences beyond the genomes (Supplementary Figure 17). However, the plastome features of both species reveal a significant variation in the patterns of IR-SC junction: *C. cathayensis* belongs to Pattern I with *C. kweichowensis*, but *C. dabieshanensis* belongs to Pattern II with *C. hunanensis* (Figure 2A). The phylogenetic topology built from the complete plastomes also supports that *C. dabieshanensis* has the close matrilineal relationship to C. hunanensis but not to *C. cathayensis* (Figure 4). Although taxonomists distinguished *C.* (*Annamocarya*) *sinensis* from the genus *Carya* by its distinctive taxonomic, botanical, and horticultural characteristics, high throughout genome-wide sequencing technology has provided a credible molecular phylogeny to distinguish plant species (Favre et al., 2020). The plastomes-based phylogenetic tree here also provided solid support that *C.* (*Annamocarya*) *sinensis* is a

member of EA *Carya* (Figure 4). From the plastome phylogenetic tree, *C. glabra* and *C. ovalis* were separated in the NA *Carya* subclade and this is also confirmed by their plastome features, the patterns of IR-SC junction variations (Figures 2A, 4), as well as their morphological differences (Grauke, 2003). As such, the matriarchal-originating plastomes provide an effective tool for taxonomic placement of the outcrossing plant species, at least at genus level.

New insight into the inconsistency between phylogenetic topologies of *Carya*

As mentioned above, maternal phylogenetic analyses based on whole plastomes and partial plastome regions, except for the IR regions, recovered the same topologies and provided substantial support of the six genera of Juglandoideae and two subclades of the genus *Carya* (Figure 4; Supplementary Figures 10-16). These results were in consistent with our early studies based on nuclear genome data and a few organelle and nuclear gene markers (Zhang et al., 2013; Huang et al., 2019). However, multiple significant inconsistencies between species of EA or NA subclades have been found in the maternal phylogenetic topologies inferred from different plastome regions and the nuclear datasets. These discordances of phylogenetic trees have been reported in numerous studies of plants in the North Hemisphere, including several close relative families of Juglandaceae, such as Betulaceae and Fagaceae in Fagales (Chan and Ragan, 2013; Lemmon and Lemmon, 2013; Zwickl et al., 2014; Stenz et al., 2015; Yang et al., 2019; Yang et al., 2021; Zhou et al., 2022). The inconsistencIES between the phylogenetic relationships based on plastomes and nuclear genes are believed to be based upon hybridization and introgression, incomplete lineage sorting, and/or ancient chloroplast capture, while the gene flow caused by hybridization and introgression, as well as plastid capture between heterocompatible species, is considered to be the source of phylogenetic inconsistencies inferred from different datasets of plastomes (Suh et al., 2015; Yang et al., 2019; Lovell et al., 2021; Yang et al., 2021; Zhou et al., 2022).

Hedrick (2013) found that introgression usually involves a small number of genes or genomic regions but may be of substantial significance. Incomplete lineage sorting largely resulted from ancestral polymorphism spanning multiple speciation events and subsequent random fixation of the polymorphisms in different lineages (Oliver, 2013), and it is common in multi-locus phylogenetic datasets due to rapid diversification (Wang et al., 2022). The entire plastome functions as a single linked locus, with different haplotypes retained through ancient clado-genic events like the alleles of a single nuclear locus (Folk et al., 2017; Lee-Yaw et al., 2019). Natural hybridization and introgression are very common in numerous wind-pollinated plant species, such as Betulaceae (Yang et al., 2019). In such species, plastid DNA is maternally inherited from ovum, whereas nuclear DNA is transmitted parentally through both pollen and ovum. Ancient and recent hybridization and introgression can result in rapid introgression of maternally inherited plastomes and the concerted evolution of the nuclear genes toward the introgressive species, especially for the geographically closely related species, and finally distort the phylogenetic relationships of introgressive species. In *Carya*, hybridization has been well studied and introgression signatures between species were evident in previous and recent reports (Lovell et al., 2021; Wang et al., 2022). In addition to the incomplete lineage sorting and ancient plastid capture, the obvious incongruencies within subclades of *Carya* between plastome- and nuclear marker-based topologies in our analysis could partially result from introgressive hybridization. Introgressive hybridization could also account for the phylogenetic incongruencies between varieties of non-monopoly *Carya* species such as *C. illinoinensis* and *C. pallida*.

'87MX3-2.11' (*C. illinoinensis*), a local seedling collection from an autochthonous tree growing near Zaachilla, Oaxaca in Mexico (Wang et al., 2020), was clustered with *C. palmeri*, which is only distributed in the tropical region in the NA *Carya* subclade (Figures 2B, 4; Supplementary Figure 16). Meanwhile, two controlled-cross cultivars of *C. illinoinensis*, 'Lakota' and 'Sioux', made in Brownwood, Texas, in 1964 and 1943, respectively (Grauke and Thompson, 1996) were grouped together (Figure 4). 'Schley' is the maternal parent of 'Sioux', while 'Lakota' was from a cross of 'Mahan' × 'Major', and 'Mahan' is a progeny of 'Schley' (Thompson and Young, 1985; Grauke et al., 2015). Therefore, it is easy to understand the sister relationships between these two cultivars (Figure 4). Meanwhile, the sister group of 'Lakota' and 'Sioux' demonstrates a close maternal phylogenetic relationship with the widely distributed species *C. tomentosa* (Figures 2B, 4; Supplementary Figure 17). Our plastome-based phylogenetic analyses also suggested that it is likely that the *C. illinoinensis* '87MXx3-2.11' share a common matrilineal ancestor with the tropical-originated *C. palmeri*, and 'Lakota' and 'Sioux' share a common female ancestor with *C. tomentosa* because of their overlapping and/or adjacent distribution (Figure 2B; Supplementary Figure 17). In addition, two *C. pallida* landraces 'Wen 12998' and 'NY1' collected from different locations of the distribution have close maternal phylogenetic relationships with *C. ovata* and *C. ovalis*, respectively. This pattern also supports the model of "the maternal phylogenetic relationship of intra-genus outcrossing species is determined by geographical distribution, and the geographically adjacent species shares a common maternal ancestor adjacent geographical distribution sharing common maternal ancestor" like the relationships among the cultivars in *C. illinoinensis* (Figure 4). Different varieties within the same species clustered in different phylogenetic positions may be caused by their multiple matrilineal origins, for reasons such as overlapping and/or adjacent distribution regions, the

hybridizability of interspecies of *Carya* species, or diverse plastome structure features, or wind pollination (Wang et al., 2022). These results may partially account for the phylogenetic incongruencies between species in EA or NA subclades inferred from partial or complete plastome datasets or nuclear data. The results generated here indicate that further broader population-wide sampling and their plastome assemblies will be a powerful approach in tracking the matrilineal historical origin of a species, especially for outcrossing species.

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: https://www.ncbi.nlm.nih.gov/genbank/, MW410235 MW410236 MW368388 ON568300 MW410227 MW255965 MW368387 MW410228 MW410229 MW410230 MW298527 MW410238 MW410237 MW410231 MW410232 MW440674 MW410233 MW410234 MW421595.

Author contributions

LX: Conceptualization. JX, SL, WZ and JW: Investigation, Resources. GX: Resources. JX and SL: Validation. HG, JH and YY: Data Curation. LX, JX, JZ and KW: Data Curation, Visualization. LX, JX and WZ: Writing Original Draft. LX, JX, WZ and XW: Writing, Review and Editing. LX, JZ: Supervision. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by grants from the Chinese Ministry of Science and Technology, China (grant no. 2018YFD1000604), the Natural Science Foundation of Zhejiang Province, China (grant no. Z20C160001), Research and Development Fund of Zhejiang A&F University (grant no. 2018FR002) and the State Key Laboratory of Subtropical Silviculture in Zhejiang A&F University, China (grants no. ZY20180202 and KF201905).

Acknowledgments

We are grateful to Dr. Zhiduan Chen at Institute of Botany, the Chinese Academy of Sciences, China, for his help and valuable suggestions during the manuscript preparation. We also thank to Drs. Dayong Zhang and Weining Bai at State Key Laboratory of Earth Surface Processes and Resource Ecology and Ministry of Education Key Laboratory for Biodiversity Science and Ecological Engineering, College of Life Sciences, Beijing Normal University, Beijing, China, for the sharing the resequencing data of *C. poilanei*.

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.2022.990064/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, 3030–3031. doi: 10.1093/bioinformatics/bty220

Bock, R. (2007). *Cell and molecular biology of plastids* (Berlin: Springer).

Chang, R. H., and Lu, A. M. (1979). A study of the genus *carya* nutt. in china. *J. Syst. Evol.* 17, 40–44.

Chan, C. X., and Ragan, M. A. (2013). Next-generation phylogenomics. *Biol. Direct* 8, 1–6. doi: 10.1186/1745-6150-8-3.

Chen, Y. X., Chen, Y. S., Shi, C. M., Huang, Z. B., Zhang, Y., Li,S. K., et al. (2018). SOAPnuke: A MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Oxford Open* 7, 1–6. doi: 10.1093/gigascience/gix120

Chevalier, A. (1941). Variabilité et Hybridité chez les Noyers. Notes sur des Juglans peu connus, sur l'Annamocarya et un Carya d'Indochine. *J. D'agriculture Traditionnelle Et Botanique Appliquée* 21, 477–509. doi: 10.3406/jatba.1941.1646

Clegg, M. T., Gaut, B. S., Learn, G. H., and Morton, A. B. R. (1994). Rates and patterns of chloroplast DNA evolution. *Proc. Natl. Acad. Sci. U. S. A.* 91, 6795–6801. doi: 10.1073/pnas.91.15.6795

Dierckxsens, N., Mardulyn, P., and Smits, G. (2016). NOVOPlasty: *De novo* assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* 45, e18. doi: 10.1093/nar/gkw955

Dong, W., Liu, Y., Li, E., Xu, C., Sun, J., Zhou, S., et al. (2022). Phylogenomics and biogeography of *Catalpa* (Bignoniaceae) reveal incomplete lineage sorting and three dispersal events. *Mol. Phyloget. Evol.* 166, 107330. doi: 10.1016/j.ympev.2021.107330

Dong, W., Xu, C., Li, C., Sun, J., Zuo, Y., Shi,S., et al. (2015). *Ycf1*, the most promising plastid DNA barcode of land plants. *Sci. Rep.* 5, 8348. doi: 10.1038/srep08348

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

Dugas, D. V., Hernandez, D., Koenen, E. J. M., 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

Favre, F., Jourda, C., Besse, P., and Charron, C. (2020). *Genotyping-by-Sequencing technology in plant taxonomy and phylogeny*. molecular plant taxonomy: Ile de la réunion. 2222, 167–178. doi: 10.1007/978-1-0716-0997-2_10

Feng, X. J., Yuan, X. Y., Sun, Y. W., Hu, Y. H., Saman, Z., Ouyang, X. H., et al. (2018). Resources for studies of iron walnut (*Juglans sigillata*) gene expression, genetic diversity, and evolution. *Tree Genet. Genomes* 14, 1–15. doi: 10.1007/s11295-018-1263-z

Folk, R. A., Mandel, J. R., and Freudenstein, J. V. (2017). Ancestral gene flow and parallel organellar genome capture result in extreme phylogenomic discord in a lineage of angiosperms. *Syst. Biol.* 66, 320–337. doi: 10.1093/sysbio/syw083

Fredrik, R., Maxim, T., Pau, V. D. M., Daniel, L. A., and Aaron, D. (2012). MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a Large model space. *Systematic Biol.* 61, 539–542. doi: 10.2307/41515220

Gao, Q. B., Yan, L., Zhuo-Ma, G., Gornall, R. J., Wang, J. L., Liu, H. R., et al. (2017). Population genetic differentiation and taxonomy of three closely related species of *saxifraga* (Saxifragaceae) from southern tibet and the hengduan mountains. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.01325

Givnish, T. J. (2016). Convergent evolution, adaptive radiation, and species diversification in plants. *Encyclopedia Evolutionary Biol.* 1, 362–373. doi: 10.1016/B978-0-12-800049-6.00266-3

Grauke, L. J. (2003). *Nut tree culture in north America*. Eds.D. Fulbright (Northern Nut Growers Association, Inc).

Grauke, L. J., Klein, R., Grusak, M. A., and Klein, P. (2015). The forest and the trees: Applications for molecular markers in the repository and pecan breeding program. *Acta Hortic.* 1070, 109–126. doi: 10.17660/ActaHortic.2015.1070.12

Grauke, L. J., and Mendoza-Herrera, M. A. (2012). Population structure in the genus *carya*. *Acta Hortic.* 948, 143–158. doi: 10.17660/ActaHortic.2012.948.16

Grauke, L. J., Mendoza-Herrera, M. A., Miller, A. J., and Wood, B. W. (2011). Geographic patterns of genetic variation in native pecans. *Tree Genet. Genomes* 7, 917–932. doi: 10.1007/s11295-011-0384-4

Grauke, L. J., and Thompson, T. E. (1996). Variability in pecan flowering. *Fruit Varieties J.* 50, 140–150.

Grauke, L. J., Wood, B. W., and Harris, M. K. (2016). Crop vulnerability: *Carya*. *HortScience* 51, 653–663. doi: 10.21273/HORTSCI.51.6.653

Grauke, L. J., Wood, B. W., and Payne, J. A. (1991). "Genetic resources of carya," in *Vietnam And china. 82nd annual report of the northern nut growers association*, vol. 82. , 80–87.

Hall, T. A. (1999). BioEdit: A user-friendly biological sequence alignment editor and analysis program for windows 95/98/NT. *Nuclc Acids Symposium Ser.* 41, 95–98. doi: 10.1021/bk-1999-0734.ch008

Hedrick, P. W. (2013). Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Mol. Ecol.* 22 (18), 4606–4618. doi: 10.1111/mec.12415

Heiges, S. B. (1896). *Nut culture in the united states: Embracing native and introduced species* (USDA Pomology Div., Govt. Printing Office: Washington D.C).

Hester, T. R. (1983). Late paleo-Indian occupations at baker cave, southwestern texas. *Bull. Tex. Arch. Soc.* 53, 101–119.

Huang, Y. J., Xiao, L. H., Zhang, Z. R., Zhang, R., Wang, Z. J., Huang, C. Y., et al. (2019). The genomes of pecan and Chinese hickory provide insights into *Carya* evolution and nut nutrition. *GigaScience* 8, 1–17. doi: 10.1093/gigascience/giz036

Hu, H., Hu, Q., Al-Shehbaz, I. A., Luo, X., Zeng, T., Guo, X., et al. (2016 1826). Species delimitation and interspecific relationships of the genus *Orychophragmus* (Brassicaceae) inferred from whole chloroplast genomes. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.01826

Hu, Y. H., Woeste, K. E., and Zhao, P. (2017 1955). Completion of the chloroplast genomes of five chinese *juglans* and their contribution to chloroplast phylogeny. *Front. Plant* 7. doi: 10.3389/fpls.2016.01955

Itaya, H., Oshita, K., Arakawa, K., and Tomita, M. (2013). GEMBASSY: An EMBOSS associated software package for comprehensive genome analyses. *Source Code Biology&Medicine* 8, 17. doi: 10.1186/1751-0473-8-17

Jiao, Y., Jia, H. M., Li, X. W., and Chai, M. L. (2012). Development of simple sequence repeat (SSR) markers from a genome survey of Chinese bayberry (*Myrica rubra*). *BMC Genomics* 13, 201. doi: 10.1186/1471-2164-13-201

Julio, R., Albert, F. M., Carlos, S. J., Sara, G. R., Pablo, L., 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

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

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

Kozlowski, G., Sébastien, B., and Song, Y. G. (2018). *Wingnuts (Pterocarya) & walnut family. relict trees: Linking the past, present and future* (Natural History Museum Fribourg: Switzerland). Relict trees: linking the past, present and future.Natural History Museum Fribourg, Switzerland.

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

Lee, S., Weon, S., Lee, S., and Kang, C. (2010). Relative codon adaptation index, a sensitive measure of codon usage bias. *Evolutionary Bioinf. Online* 6, 47–55. doi: 10.4137/EBO.S4608

Lee-Yaw, J. A., Grassa, C. J., Joly, S., Andrew, R. L., and Rieseberg, L. H. (2019). An evaluation of alternative explanations for widespread cytonuclear discordance in annual sunflowers (*Helianthus*). *New Phytol.* 221 (1), 515–526. doi: 10.1111/nph.15386

Lemmon, E. M., and Lemmon, A. R. (2013). High-throughput genomic data in systematics and phylogenetics. *Annu. Rev. Ecol. Evol. Systematics* 44, 99–121. doi: 10.1111/nph.15386

Leroy, J. F. (1950). Note sur les noyers (*Carya et annamocarya*) sauvages d'Indochine. *Rev. Internationale Botanique Appliquée Et D'agriculture Tropicale* 30, 425–428. doi: 10.3406/jatba.1950.6726

Leroy, J. F. (1955). Etude sur les juglandaceae. *Mem. Mus. Hist. Nat. (Paris)II B [Bot.]* 6, 246. Available at: https://www.biodiversitylibrary.org/item/266009#page/1/mode/1up.

Li, Y. X., Li, Z. H., Schuitman, A., Chase, M. W., and Jin, X. H. (2019). Phylogenomics of orchidaceae based on plastid and mitochondrial genomes. *Mol. Phylogenet. Evol.* 139, 106540. doi: 10.1016/j.ympev.2019.106540

Li, H., Liu, B., Davis, C. C., and Yang, Y. (2020). Plastome phylogenomics, systematics, and divergence time estimation of the *Beilschmiedia* group (Lauraceae). *Mol. Phylogenet. Evol.* 151, 106901. doi: 10.1016/j.ympev.2020.106901

Little, E. L. (1969). Two varietal transfers in *Carya* (Hickory). *Phytologia* 19, 186–190.

Liu, M. C., and Li, Z. J. (1984). A new species of carya from China. *J. Zhejiang F Univ.* 1, 49–51.

Liu, J., Lindstrom, A. J., and Gong, X. (2022). Towards the plastome evolution and phylogeny of *Cycas* l. (Cycadaceae): molecular-morphology discordance and gene tree space analysis. *Front. Plant Sci.* 22, 116. doi: 10.1186/s12870-022-03491-2

Lovell, J. T., Bentley, N. B., Bhattarai, G., Jenkins, J. W., Sreedasyam, A., Alarcon, Y., et al. (2021) 4125. Four chromosome scale genomes and a pan-genome annotation to accelerate pecan tree breeding. *Nat. Commun.* 12, 4125. doi: 10.1038/s41467-021-24328-w

Lu, A. M., Stone, D. E., and Grauke, L. J. (1999). *Juglandaceae, in flora of china. China/Missouri botanical garden*: Beijing.

Lu, Q. X., Ye, W. Q., Lu, R. S., Xu, W. Q., and Qiu, Y. X. (2018). Phylogenomic and comparative analyses of complete plastomes of *croomia* and *stemona* (Stemonaceae). *Int. J. Mol. Sci.* 19, 2383. doi: 10.3390/ijms19082383

Luo, J., Chen, J., Guo, W., Yang, Z., Lim, K.-J., and Wang, Z. (2021). Reassessment of *Annamocarya sinesis* (*Carya sinensis*) taxonomy through concatenation and coalescence phylogenetic analysis. *Plants* 11, 52. doi: 10.3390/plants11010052

Mader, M., Pakull, B., Blanc, J. C., Drewes, M. P., Zoéwindé, B., Bernd, D., et al. (2018). Complete chloroplast genome sequences of four meliaceae species and comparative analyses. *Int. J. Mol. Sci.* 19, 701. doi: 10.3390/ijms19030701

Mai, D. H. (1981). Der formenkreis der Vietnam-nuß (*Carya poilanei* (Chev.) Leroy) in Europa. *Feddes Repert* 92, 339–385. doi: 10.1002/fedr.4910920502

Manning, W. E. (1950). A key to the hickories north of Virginia with notes on the two pignuts *Carya glabra* and *C. ovalis*. *Rhodora* 52, 188–199. doi: 10.2307/23304073

Manning, E. W. (1963). Hickories reported in India and Laos with other notes on *Carya* in Asia. *Brittonia* 15, 123–125. doi: 10.2307/2805397

Manning, W. E. (1978). The classification within the juglandaceae. *Ann. Missouri Botanical Garden* 65, 1058–1087. doi: 10.2307/2398782

Manning, W. E., and Hjelmqvist, H. (1951). *Annamocarya rhampocarya* and *Carya sinensis*. *Bot. Notiser* 4, 319–330.

Manos, P. S., Soltis, P. S., Soltis, D. E., and Manchester, S. R. (2007). Phylogeny of extant and fossil juglandaceae inferred from the integration of molecular and morphological data sets. *Systematic Biol.* 56, 412–430. doi: 10.1080/10635150701408523

Manos, P. S., and Stone, D. E. (2001). Evolution, phylogeny, and systematics of the juglandaceae. *Ann. Missouri Botanical Garden* 88, 231–269. doi: 10.2307/2666226

Meng, X. X., Xian, Y. F., Li, X., Dong, Z., Shi, Y. H., Wu, M. L., et al. (2018). Complete chloroplast genomes from sanguisorba: Identity and variation among four species. *Molecules* 23, 2137. doi: 10.3390/molecules23092137

Michael, T., Pascal, L., Tommaso, P., Ulbricht-Jones, E. S., Axel, F., Ralph, B., et al. (2017). GeSeq – versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* W1. doi: 10.1093/nar/gkx391

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

Mu, X. Y., Tong, L., Sun, M., Zhu, Y. X., Wen, J., Lin, Q. W., et al. (2020). Phylogeny and divergence time estimation of the walnut family (Juglandaceae) based on nuclear RAD-seq and chloroplast genome data. *Mol. Phylogenet. Evol.* 147, 106802. doi: 10.1016/j.ympev.2020.106802

Ogoma, C. A., Liu, J., Stull, G. W., Wambulwa, M. C., Oyebanji, O., Miline, R. I., et al. (2022). Deep insights into the plastome evolution and phylogenetic relationships of the tribe *Urticeae* (Family urticaceae). *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.870949

Oliver, J. C. (2013). Microevolutionary processes generate phylogenomic discordance at ancient divergences. *Evolution* 67 (6), 1823–1830. doi: 10.1111/evo.12047

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. J. (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

Peng, S. B., Yang, G. Y., Liu, C. B., Yu, Z. D., and Zhai, M. Z. (2017). The complete chloroplast genome of the *Juglans regia* (Juglandales: Julandaceae). *Mitochondrial DNA* 28, 407–408. doi: 10.3109/19401736.2015.1127367

Peter, R., Ian, L., and Alan, B. (2000). EMBOSS: The european molecular biology open software suite. *Trends Genet.* 16, 276–277. doi: 10.1016/s0168-9525(00)02024-2

Pham, K. K., Hipp, A. L., and Manos, P. S. (2017). And cronn, rA time and a place for everything: phylogentic history and geography as joint predictors of oak plastome phylogeny. *C. Genome* 60, 720–732. doi: 10.1139/gen-2016-0191

Raubeson, L. A., and Jansen, R. K. (2005). "Chloroplast genomes of plants," in *Plant diversity and evolution: Genotypic and phenotypic variation in higher plants*.

Raubeson, L. A., Peery, R., Chumley, T. W., Dziubek, C., Fourcade, H. M., Boore, J. L., et al. (2007). Comparative chloroplast genomics: Analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. *BMC Genomics* 8, 174. doi: 10.1186/1471-2164-8-174

Raven, P. H. (1972). Plant species disjunctions: a summary. *Ann. Mo. Bot. Gard.* 59, 234–246. doi: 10.2307/2394756

Scott, R. A. (1953). Status of an asiatic member of the juglandaceae regarded as a 'Living fossil'. *Am. J. Bot.* 40, 666–669. doi: 10.1002/j.1537-2197.1953.tb06538.x

Sebastian, B., Thomas, T., Münch, T., Uwe, S., and Martin, M. (2017). MISA-web: A web server for microsatellite prediction. *Bioinf. (Oxford England)* 33, 2583–2585. doi: 10.1093/bioinformatics/btx198

Song, F., Li, T., Burgess, K. S., Feng, Y., and Ge, X.-J. (2020). Complete plastome sequencing resolves taxonomic relationships among species of *Calligonum* l. (Polygonaceae) in China. *BMC Plant Biol.* 20, 261. doi: 10.1186/s12870-020-02466-5

Stenz, N. W., Larget, B., Baum, D. A., and An, C. C. (2015). Exploring tree-like and non-tree-like patterns using genome sequences: An example using the inbreeding plant species *Arabidopsis thaliana* (L.) heynh. *Systematic Biol.* 64, 809–823. doi: 10.1093/sysbio/syv039

Stephan, G., Pascal, L., and Ralph, B. (2019). OrganellarGenomeDRAW (OGDRAW) version 1.3.1: Expanded toolkit for the graphical visualization of organellar genomes. *Nuclc Acids Res.* 47, W59–W64. doi: 10.1093/nar/gkz238

Suh, A., Smeds, L., and Ellegren, H. (2015). The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. *PloS Biol.* 13, e1002224. doi: 10.1371/journal.pbio.1002224

Thach, R. E., Dewey, K. F., Brown, J. C., and Doty, P. (1966). Formylmethionine codon AUG as an initiator of polypeptide synthesis. *Science* 153, 416–418. doi: 10.1126/science.153.3734.416

Thompson, T. E., Grauke, L. J., and Reid, W. (2008). 'Lakota' pecan. *Hortscience Publ. Am. Soc. Hortic. Sci.* 43. doi: 10.21273/HORTSCI.43.1.250

Thompson, T. E., and Young, F. (1985). *Pecan cultivars: Past and present*.

Thompson, T. E., and Grauke, L. J. (1991). Pecans and other hickories (*Carya*). *Acta Hortic* 290, 839–906. doi: 10.17660/ActaHortic.1991.290.19

Tiffney, B. H. (1985). The Eocene north Atlantic land bridge: its importance in tertiary and modern phytogeography of the northern hemisphere. *J. Arnold Arboretum* 66, 243–273.

Tu, X.-D., Liu, D.-K., Xu, S.-W., Zhou, C.-Y., Gao, X.-Y., Zeng, M.-Y., et al. (2021). Plastid phylogenomics improves resolution of phylogenetic relationship in the cheirostylis and *Goodyera* clades of *Goodyerinae* (Orchidoideae, rchidaceae). *Mol. Phylogen. Evol.* 164, 107269. doi: 10.1016/j.ympev.2021.107269

Wang, R. J., Cheng, C. L., Chang, C. C., Wu, C. L., Su, T. M., Chaw, S. M., et al. (2008). Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. *BMC Evolutionary Biol.* 8, 36–51. doi: 10.1186/1471-2148-8-36

Wang, X.-H., Moore, M. J., Barrett, R. L., Landrein, S., Sakaguchi, S., Maki, M., et al. (2020). Plastome phylogenomic insights into the sino-Japanese biogeography of *Diabelia* (Caprifoliaceae). *J. System. Evol.* 58 (6), 972–987. doi: 10.1111/jse.12560

Wang, X. W., Rhein, H. S., Jenkins, J., Schmutz, J., and Randall, J. J. (2020). Chloroplast genome sequences of *Carya illinoinensis* from two distinct geographic populations. *Tree Genet. Genomes* 16, 2–14. doi: 10.1007/s11295-020-01436-0

Wang, Y., Ruhsam, M., Milne, R., Graham, S. W., Li, J., Tao, T., et al. (2022). Incomplete lineage sorting and local extinction shaped the complex evolutionary history of the paleogene relict conifer genus, *Chamaecyparis* (Cupressaceae). *Mol. Phylogen. Evol.* 172, 107485. doi: 10.1016/j.ympev.2022.107485

Wang, X. M., Zhou, T., Bai, G. Q., and Zhao, Y. M. (2018). Complete chloroplast genome sequence of *Fagopyrum* dibotrys: Genome features, comparative analysis and phylogenetic relationships. *Sci. Rep-Uk* 8, 12379. doi: 10.1038/s41598-018-30398-6

Wei, R., Yan, Y.-H., Harris, A. J., Kang, J.-S., Shen, H., Xiang, Q.-P., et al. (2017). Plastid phylogenomics resolve deep relationships among eupolypod II ferns with rapid radiation and rate heterogeneity. *Genome Biol. Evol.* 9 (6), 1646–1657. doi: 10.1093/gbe/evx1075

Wen, J. (1999). Evolution of eastern Asian and eastern north American disjunct distributions in flowering plants. *Annu. Rev. Ecol. Syst.* 30, 421–455. doi: 10.1146/annurev.ecolsys.30.1.421

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.* 31, 645–659. doi: 10.1093/molbev/mst257

Xiao., L., Yu, M., Zhang, Y., Hu, J., Wang, J., Guo, H., et al. (2021). Chromosome-scale assembly reveals the asymmetric paleo-subgenomes evolution and targets for accelerating fungal resistance breeding in nut crop, pecan. *Plant Commun*. doi: 10.1016/j.xplc.2021.100247

Xu, C., Dong, W., Li, W., Lu, Y., Xie, X., Jin, X., et al. (2017). Comparative analysis of six Lagerstroemia complete chloroplast genomes. *Front. Plant Sci.* 8, 15. doi: 10.3389/fpls.2017.00015

Yan, M., Fritsch, P. W., Moore, M. J., Feng, T., Meng, A., Yang, J., et al. (2018). Plastid phylogenomics resolves infrafamilial relationships of the styracaceae and sheds light on the backbone relationships of the ericales. *Mol. Phylogenet. Evol.* 121, 198–211.

Yang, Y.-Y., Qu, X.-J., Zhang, R., Stull, G. W., and Yi, T.-S. (2021). Plastid phylogenomic analyses of fagales reveal signatures of conflict and ancient chloroplast capture. *Mol. Phylogenet. Evol.* 163, 107232. doi: 10.1016/j.ympev.2021.107232

Yang, X. Y., Wang, Z. F., Luo, W. C., Guo, X. Y., Zhang, C. H., Liu, J. Q., et al. (2019). Plastomes of betulaceae and phylogenetic implications. *J. Systematics Evol.* 57, 508–518. doi: 10.1111/jse.12479

Yang, A. H., Zhang, J. J., Yao, X. H., and Huang, H. W. (2011). Chloroplast microsatellite markers in *Liriodendron tulipifera* (Magnoliaceae) and cross-species amplification in *L. chinense*. *Am. J. Bot.* 98, e123–e126. doi: 10.3732/ajb.1000532

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

Yao, G., Jin, J. J., Li, H. T., Yang, J. B., Shiva, M. V., Croley, M., et al. (2019). Plastid phylogenomic insights into the evolution of caryophyllales. *Mol. Phylogenet. Evol.* 134, 74–86. doi: 10.1016/j.ympev.2018.12.023

Ye, L. J., Fu, C. N., Wang, Y. H., Liu, J., and Gao, L. M. (2018). Characterization of the complete plastid genome of a Chinese endemic species *Carya kweichowensis*. *Mitochondrial DNA Part B* 3, 492–493. doi: 10.1080/23802359.2018.1464414

Zhai, D. C., Yao, Q., Cao, X. F., Hao, Q. Q., and Bai, X. H. (2019). Complete chloroplast genome of the wild-type hickory *Carya cathayensis*. *Mitochondrial DNA Part B* 4, 1457–1458. doi: 10.1080/23802359.2019.1598815

Zhang, W. P., Bai, W. N., and Zhang, D. Y. (2022). The rediscovery of *Carya poilanei* (Juglandaceae) after 63 years reveals a new record from China. *PhytoKeys* 188, 73–82. doi: 10.3897/phytokeys.188.77242

Zhang, J. B., Li, R. Q., Xiang, X. G., Manchester, S. R., Li, L., Wang, W., et al. (2013). Integrated fossil and molecular data reveal the biogeographic diversification of the eastern Asian-Eastern north american disjunct hickory genus (*Carya* nutt.). *PloS One* 8, e70449. doi: 10.1371/journal.pone.0070449

Zhang, Q. Y., Ree, R. H., Salamin, N., Xing, Y. W., and Silvestro, D. (2021). Fossil-informed models reveal a boreotropical origin and divergent evolutionary trajectories in the walnut family (Juglandaceae). *Systematic Biol.* 0, 1–17. doi: 10.1093/sysbio/syab030

Zhang, B. W., Xu, L. L., Li, N., Yan, P. C., Jiang, X. H., Woeste, K. E., et al. (2019). Phylogenomics reveals an ancient hybrid origin of the persian walnut. *Mol. Biol. Evol.* doi: 10.1093/molbev/msz112

Zhao, P., Zhou, H. J., Potter, D., Hu, Y. H., Feng, X. J., Dang, M., et al. (2018). Population genetics, phylogenomics and hybrid speciation of juglans in China determined from whole chloroplast genomes, transcriptomes, and genotyping-by-sequencing (GBS). *Mol. Phylogenet. Evol.* 126, 250–265. doi: 10.1016/j.ympev.2018.04.014

Zhou, H., Hu, Y., Ebrahimi, A., Liu, P., and Zhao, P. (2021). Whole genome based insights into the phylogeny and evolution of the juglandaceae. *BMC Ecol. Evol.* 21, 191. doi: 10.21203/rs.3.rs-495294/v1

Zhou, B.-F., Yuan, S., Crowl, A. A., Liang, Y.-Y., Shi, Y.-Y., Chen, X.-Y., et al. (2022 1320). Phylogenomic analyses highlight innovation and introgression in the continental radiations of fagaceae across the northern hemisphere. *Nat. Commun.* 13, 1320. doi: 10.1038/s41467-022-28917-1

Zwickl, D. J., Stein, J. C., Wing, R. A., Ware, D., and Sanderson, M. J. (2014). Disentangling methodological and biological sources of gene tree discordance on *Oryza* (Poaceae) chromosome 3. *Systematic Biol.* 63, 645–659. doi: 10.1093/sysbio/syu027

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 10 November 2022
DOI 10.3389/fpls.2022.1030647

OPEN ACCESS

EDITED BY
Wenpan Dong,
Beijing Forestry University, China

REVIEWED BY
Yiheng Wang,
State Key Laboratory of Dao-di Herbs, China Academy of Chinese Medical Sciences, China
Guanglong Hu,
Beijing Academy of Agricultural and Forestry Sciences, China

*CORRESPONDENCE
Xiao-Man Xie
xxm529@126.com
Wen-Qing Li
350127263@qq.com

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

RECEIVED 29 August 2022
ACCEPTED 21 October 2022
PUBLISHED 10 November 2022

CITATION
Han B, Zhang M-J, Xian Y, Xu H, Cui C-C, Liu D, Wang L, Li D-Z, Li W-Q and Xie X-M (2022) Variations in genetic diversity in cultivated *Pistacia chinensis*. *Front. Plant Sci.* 13:1030647. doi: 10.3389/fpls.2022.1030647

Variations in genetic diversity in cultivated *Pistacia chinensis*

Biao Han[1], Ming-Jia Zhang[2], Yang Xian[1], Hui Xu[1], Cheng-Cheng Cui[1], Dan Liu[1], Lei Wang[1], De-Zhu Li[3], Wen-Qing Li[1]* and Xiao-Man Xie[1]*

[1]Key Laboratory of State Forestry and Grassland Administration Conservation and Utilization of Warm Temperate Zone Forest and Grass Germplasm Resources, Shandong Provincial Center of Forest and Grass Germplasm Resources, Ji'nan, Shandong, China, [2]College of Forestry, Shandong Agricultural University, Tai'an, Shandong, China, [3]Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan, China

Identification of the evolution history and genetic diversity of a species is important in the utilization of novel genetic variation in this species, as well as for its conservation. *Pistacia chinensis* is an important biodiesel tree crop in China, due to the high oil content of its fruit. The aim of this study was to uncover the genetic structure of *P. chinensis* and to investigate the influence of intraspecific gene flow on the process of domestication and the diversification of varieties. We investigated the genetic structure of *P. chinensis*, as well as evolution and introgression in the subpopulations, through analysis of the plastid and nuclear genomes of 39 P. *chinensis* individuals from across China. High levels of variation were detected in the *P. chinensis* plastome, and 460 intraspecific polymorphic sites, 104 indels and three small inversions were identified. Phylogenetic analysis and population structure using the plastome dataset supported five clades of *P. chinensis*. Population structure analysis based on the nuclear SNPs showed two groups, clearly clustered together, and more than a third of the total individuals were classified as hybrids. Discordance between the plastid and nuclear genomes suggested that hybridization events may have occurred between highly divergent samples in the *P. chinensis* subclades. Most of the species in the *P. chinensis* subclade diverged between the late Miocene and the mid-Pliocene. The processes of domestication and cultivation have decreased the genetic diversity of *P. chinensis*. The extensive variability and structuring of the *P. chinensis* plastid together with the nuclear genomic variation detected in this study suggests that much unexploited genetic diversity is available for improvement in this recently domesticated species.

KEYWORDS

discordance, genetic diversity, nuclear SNPs, *Pistacia chinensis*, plastome

Introduction

The genus *Pistacia* (Anacardiaceae) consists of at least 11 species (Parfitt and Badenes, 1997; Xie et al., 2014). The species are trees or shrubs and are dioecious, with female and male flowers on separate trees, and the fruit is a monocarpic drupe. Chinese pistache tree (*Pistacia chinensis* Bunge) is a small, wind-pollinated tree species with apetalous flowers, and is widely distributed throughout China owing to its strong adaptability to poor habitat and adverse conditions. This tree has potential as a biodiesel tree species in China due to the high oil content of its fruit (Li et al., 2010). The oil content in the seed is typically higher than 40% and the sixteen alkyl value of biodiesel derived from the seeds is generally up to 51.3 (Wang et al., 2012). *Pistacia chinensis* has been used as a landscape tree and as a vegetable, and is also used as rootstock for *P. vera*, because it is strongly adaptable and resistant to adverse conditions (Tang et al., 2012). Additionally, *P. chinensis* is also used in Chinese traditional medicine to relieve dysentery, inflammatory swelling, psoriasis and rheumatism (Tang et al., 2012).

Deterministic or stochastic forces, such as domestication or genetic drift, may decrease genetic diversity at different levels of biological organization, for example at the individual, population or species level. Evolutionary forces such as dispersal, hybridization or introgression can lead to decreases or increases in divergence among difference subpopulations, obscuring the origins of domestication and mixing the genetic variations. *P. chinensis* is a recently domesticated species, and in this important biodiesel tree species, understanding the genetic diversity of the wild germplasm is essential in order to prioritize conservation of novel wild germplasm that may be useful in the future, and to guide the introduction of novel genetic diversity into selective breeding populations.

Genetic diversity in *P. chinensis* has previously been studied using several markers, including SSR (Wu et al., 2010; Lu et al., 2019; Cheng et al., 2022), random amplified microsatellite polymorphism (RAMP), random amplified polymorphic DNA (RAPD) and amplified fragment length polymorphism (AFLP) (Katsiotis et al., 2003). Parfitt and Badenes and Xie et al. (Parfitt and Badenes, 1997; Xie et al., 2014) used plastid and nuclear genome sequences in phylogenetic and biogeographic analyses of *Pistacia*, however, to date, few studies, if any, have examined extensive genomic SNP data (including the plastid and nuclear genomes) for population analyses. The advantages of DNA sequence markers include their high reproducibility, increasing the chances and the abilities of detecting genetic diversity (Kumar et al., 2009).

With advances in sequencing methods, genomic data in particular are a popular in the evaluation of population genetics (Fu et al., 2022; Karbstein et al., 2022). This has led to scientists focusing on the nuclear genome and paying little attention to the plastome, which is considered to have lower divergence within species. However, recently, certain evolutionary studies have been conducted at the intraspecies level based on plastomes, providing further insight into the biogeographical structure and extensive genetic variation at the population level (Perdereau et al., 2017; Magdy et al., 2019; Migliore et al., 2019; Mohamoud et al., 2019). Plastids, with their highly conserved maternally inherited genomes, show a clear geographical structure (Hohmann et al., 2018; Xue et al., 2021), and are therefore useful in phylogeographical studies. Therefore, combining plastid and nuclear genome sequences allows us to conduct comprehensive investigations into genetic diversity because the genetic information available is maternally and biparentally inherited, respectively.

In this study, we sequenced the genomes of 39 P. *chinensis* individuals from across China. Whole plastomes and nuclear SNPs were assembled and analyzed, and we then used this huge genetic variation to characterize the structure and diversity of *P. chinensis* from the differences among the individuals. We then compared the plastid and nuclear genomes and investigated possible gene flow and introgression occurring during the domestication of *P. chinensis*. Our results not only reveal evolutionary factors responsible for reshaping the genetic variation in *P. chinensis* populations, but also exemplify changes in genetic diversity during the domestication and cultivation processes.

Materials and methods

Sampling and DNA extraction

We collected a total of 39 samples of *P. chinensis* from across China and deposited them in the Shandong Provincial Center of Forest Tree Germplasm Resources, Jinan, China (Table 1). Due to the widespread cultivation of *Pistacia chinensis*, we conducted random sampling of accessions from the same area. The sampled accessions were all collected from within the natural range of this species. Fresh leaf material was dried in silica gel, and voucher specimens were deposited in the herbarium of the Shandong Provincial Center of Forest Tree Germplasm Resources. About 0.02 g of dried leaf tissue was ground using a mechanical lapping method, and total genomic DNA was extracted using a modified CTAB protocol (Li et al., 2013). DNA concentration was quantified using a Qubit 2.0 Fluorometer (Thermo Fisher Scientific, Inc., Carlsbad, CA, USA), and the size and quality of the DNA were visualized and assessed using a TAE agarose gel.

Library preparation and sequencing

A total of 700 ng DNA per sample was used for library preparation. Genomic DNA was fragmented by sonication into

TABLE 1 Sample information and the size of the chloroplast genome sequences.

Voucher	Samples	Location	LSC	IR	SSC	Total length	Genbank accession number
BJFC 00021725	Anhui	Huaning County, Huangshan City, Anhui Province	88371	26595	19057	160618	OP554543
BJTZS	Beijing	Beijing City	88371	26595	19057	160618	OP554540
202109ZPF0133	Gansu	Hui County, Longnan City, Gansu Province	88371	26595	19057	160618	OP554537
BJFC 00021710	Guangxi	Nanning City, Guangxi Province	88526	26595	19090	160806	OP554555
BJFC 00076565	Guizhou01	Congjiang County, Qiandongnan, Guizhou Province	88371	26595	19057	160618	OP554538
BJFC 00078597	Guizhou02	Rongjiang County, Qiandongnan, Guizhou Province	88366	26595	19087	160643	OP554557
202009zlj069	Hebei01	Wuan City, Handan, Hebei Province	88371	26595	19057	160618	OP554536
202009zlj070	Hebei02	Wuan City, Handan, Hebei Province	88371	26595	19057	160618	OP554535
BJFC 00081493	Henan01	Xinxiang City, Henan Province	88371	26595	19057	160618	OP554541
BJFC 00081484	Henan02	Xinxiang City, Henan Province	88387	26596	19088	160667	OP554550
BJFC 00079805	Henan03	Xinxiang City, Henan Province	88387	26596	19088	160667	OP554549
BJFC 00085509	Henan04	Xinxiang City, Henan Province	88388	26596	19088	160668	OP554552
BJFC 00100722	Jiangsu01	Nanajing City, Jiangsu Province	88371	26595	19057	160618	OP554539
202009bzg003	Jiangsu02	Haizhou District, Lianyungang City, Jiangsu Province	88371	26595	19057	160618	OP554534
202009bzg020	Jiangsu03	Haizhou District, Lianyungang City, Jiangsu Province	88371	26595	19057	160618	OP554533
202009bzg083	Jiangsu04	Lianyun District, Lianyungang City, Jiangsu Province	88546	26595	19085	160821	OP554553
202009bzg154	Jiangsu05	Pukou District, Nanjing City, Jiangsu Province	88371	26595	19057	160618	OP554532
BJFC 00021716	Jiangxi	Xunwu County, Ganzhou City, Jiangxi Province	88546	26595	19084	160820	OP554554
201909wfs049	Shandong01	Muping District, Yantai City, Shandong Province	88371	26595	19057	160618	OP554531
201910lq011	Shandong02	Changqing District, Jinan City, Shandong Province	88387	26596	19087	160666	OP554551
201910lw012	Shandong03	Huangdao District, Qingdao City, Shandong Province	88371	26595	19057	160618	OP554529
201910lw014	Shandong04	Huangdao District, Qingdao City, Shandong Province	88371	26595	19057	160618	OP554542
201910lw016	Shandong05	Huangdao District, Qingdao City, Shandong Province	88371	26595	19057	160618	OP554528
201910lw017	Shandong06	Huangdao District, Qingdao City, Shandong Province	88371	26595	19057	160618	OP554527
201910lw005	Shandong07	Huangdao District, Qingdao City, Shandong Province	88371	26595	19057	160618	OP554530
202009hb040	Shandong08	Tai'an City, Shandong Province	88387	26596	19088	160667	OP554548
202009gxj042	Shandong09	Tantai City, Shandong Province	88371	26595	19057	160618	OP554526
202009wl165	Shanxi01	Mei County, Baoji City, Shaanxi Province	88371	26595	19057	160618	OP554525
202110040028	Shanxi02	Feng County, Baoji City, Shaanxi Province	88371	26595	19057	160618	OP554521
202110170077	Shanxi03	Feng County, Baoji City, Shaanxi Province	88371	26595	19057	160618	OP554520
202110040060	Shanxi04	Feng County, Baoji City, Shaanxi Province	88405	26596	19089	160686	OP554544
202110040075	Shanxi05	Feng County, Baoji City, Shaanxi Province	88371	26595	19057	160618	OP554522
202109WL0154	Shanxi06	Loyang County, Hanzhong City, Shaanxi Province	88405	26596	19089	160686	OP554545
202109WL0099	Shanxi07	Loyang County, Hanzhong City, Shaanxi Province	88371	26595	19057	160618	OP554523
202109WL0093	Shanxi08	Loyang County, Hanzhong City, Shaanxi Province	88405	26596	19089	160686	OP554546
202109WL0064	Shanxi09	Loyang County, Hanzhong City, Shaanxi Province	88405	26596	19089	160686	OP554547
202009wl251	Shanxi10	Zhouzhi County, Xi'an City, Shaanxi Province	88371	26595	19057	160618	OP554524
BJFC 00021739	Yunnan01	Kunming City, Yunan Province	88481	26595	19083	160754	OP554556
202110110137	Zhenjiang	Lin'an District, Hangzhou, Zhejiang Province	88371	26595	19057	160618	OP554519

350 bp fragments. Sequencing libraries were generated using NEB Next® Ultra™ DNA Library Prep Kit for Illumina (NEB, USA) and was then used for sequencing. Each sample was barcoded with a unique index, and libraries were pooled. Whole-genome shotgun sequence data was paired-end sequenced (2 × 150 bp) on an Illumina HiSeq X-ten platform (Illumina, Inc., San Diego, CA, USA). Most samples yielded approximately 15 Gb of 150-bp paired-end reads, which is about 30 X depth of coverage for the genome of *Pistacia chinensis*.

Assembly and annotation of the plastome

Quality control of raw reads was conducted using Trimmomatic version 0.39 (Bolger et al., 2014) with the following options: LEADING, 20; TRAILING, 20; SLIDING WINDOW, 4:15; MIN LEN, 36; and AVG QUAL, 20. Clean reads were used to assemble the plastome of *P. chinensis* using GetOrganelle, with a range of k-mers of 75, 85, 95, and 105 (Jin

et al., 2020). Where GetOrganelle failed to assemble the complete plastome, we assembled it following the methods of Dong et al. (2022). Gene annotation of the plastome was performed with Plann (Huang and Cronk, 2015), and the published genome of *P. chinensis* (GenBank accession number: MT157378) was used as the reference sequence. The physical map of the *P. chinensis* plastome was drawn in Chloroplot (Zheng et al., 2020).

Analysis of variation in the plastome

The genome sequences from the 39 P. *chinensis* individuals were aligned using MAFFT version 7.490 (Katoh and Standley, 2013) and adjusted manually using Se-Al version 2.0 (Rambaut, 1996) to avoid alignment errors, such as polymeric repeat structures and small inversions. Nucleotide diversity, number of indels and sequence distance were used to assess sequence divergence over all the plastomes. The number of variable sites, parsimony-informative sites and sequence distances (π) were calculated using MEGA version 7.0 (Kumar et al., 2016). Nucleotide diversity and number of indels were calculated using DnaSP version 6 (Rozas et al., 2017).

Reference mapping and nuclear SNP calling

Clean reads were mapped to the pistachio (*Pistacia vera*) reference genome (Zeng et al., 2019) using the program BWA version 0.7.17 (Li and Durbin, 2010) with default settings. Potential PCR duplicates were removed using SAMtools version 1.3.1 (Li et al., 2009). Only uniquely mapped paired reads were used for the detection of SNPs. The high-quality nuclear SNPs were called through GATK version 4.2.0.0 (Heldenbrand et al., 2019) and Picard tools version 1.92 (http://broadinstitute.github.io/picard/). The SNPs were then extracted and filtered according to the following criteria: quality value ≥ 20; sites with coverage over 2; missing data less then 10%. The SNP VCF files were then merged together with VCFtools version 0.1.14.

Phylogenetic analysis of *Pistacia chinensis* individuals

The plastome sequences of the 39 sampled *P. chinensis* individuals, with that of *P. weinmaniifolia* as the outgroup, were aligned using MAFFT version 7.490 (Katoh and Standley, 2013). A phylogenetic tree based on this plastome dataset was then reconstructed using a maximum likelihood (ML) method in RAxML-NG (Kozlov et al., 2019). The best-fit model for ML analysis was found to be ModelFinder (Kalyaanamoorthy et al., 2017) based on Bayesian information criteria.

P. vera was used as the outgroup for the nuclear SNPs dataset. In order to investigate intraspecific hybridization among the individuals, we used the following dividing method to infer the phylogenetic relationships within the nuclear SNPs dataset, thereby avoiding concatenation-based ML analyses. In this method, each 100 kb of SNPs were divided into a new data matrix and used for tree reconstruction. ML trees were inferred using IQ-TREE version 2 (Minh et al., 2020) and branch support values were computed using the UFBoot method.

Population structure and PCA analysis

The plastome and nuclear SNPs datasets were used to examine the population ancestry. ADMIXTURE was used to investigate the population genetic structure of all individuals, specifying K values ranging from 1 to 10 (Alexander and Lange, 2011). The optimum number of clusters (K) was determined at the K value with the lowest cross-validation error. Principal component analysis (PCA) was also conducted to evaluate the genetic structure of *P. chinensis* using Plink (Purcell et al., 2007), and graphs were built using the ggbiplot package in R. We constructed a network using the plastome dataset. Haplotype data were analyzed in DnaSP version 6 (Rozas et al., 2017) and a TCS network was built using PopArt version 1.7 (Clement et al., 2002; Leigh and Bryant, 2015).

Analysis of intraspecific hybridization

We divided all samples into three populations with multiple individuals according to the Admixture result K = 2. TreeMix version 1.12 (Pickrell and Pritchard, 2012) was used to estimate gene flows between different populations, with blocks of 200 SNPs to account for linkage disequilibrium, and standard errors of migration rates were also calculated. The outputs were visualized in R.

Estimation and profiling of divergence time

We used the complete plastome to estimate the divergence times of the different haplotypes. This dataset included 12 haplotypes of *P. chinensis* and a further 28 species from the Anacardiaceae. Four priors were used for this analysis. The root age of the tree (crown age of Anacardiaceae) was set to 70 Ma according to the wood fossils related to the Anacardiaceae and Burseraceae, which were reported from the upper Cretaceous of Mexico. The minimum stem age of *Rhus* was calibrated as 44

Ma, according to the fruit fossils of *Rhus*, from the middle Eocene of western North America (Manchester, 1994). The other two priors were taken from the findings of Xie et al. (Xie et al., 2014): the crown age of *Rhus* was set as 33.24 Ma and the stem age of *Cotinus* was set to 37.6 Ma.

BEAST 2 (Bouckaert et al., 2014) was used to perform the divergence time analyses. A GTR model and an uncorrelated lognormal distribution relaxed molecular clock model were selected. A Markov Chain Monte Carlo (MCMC) algorithm was run for 500,000,000 generations, sampling every 50,000 generations. Convergence was assessed using Tracer version 1.6 (Rambaut et al., 2014) with effective sampling sizes (ESS) in all parameters surpassing 200. The first 10% of the trees were discarded as burn-in and the remaining trees were used to construct the Maximum Clade Credibility (MCC) tree with mean heights in TreeAnnotator.

Results

The *Pistacia chinensis* plastome and sequence variation

The plastome of *P. chinensis* ranged from 160,618 to 160,821 bp in length (Table 1) and consisted of four distinct parts, including a large single copy (LSC) region, a small single copy (SSC) region, and a pair of inverted repeats (IRA/IRB), exhibiting similar structure typical of most angiosperm species (Figure 1 and Figure S1). The LSC (between 88,366 bp and 88,546 bp) and SSC (between 19,057 bp and 19,090 bp) were separated by the two IR regions (between 26,595 bp and 26,596 bp). The overall GC content was 37.9%, and the GC content was slightly higher in the IR (42.9%) regions than in the LSC (36.0%) and SSC (32.4%) regions. The annotated *P. chinensis* plastome included 113 unique genes (79 protein-coding genes, 30 tRNA genes, and four rRNA genes), with 60 protein-coding and 22 tRNA genes in the LSC, 11 protein-coding and one tRNA genes in the SSC, and with eight protein-coding genes, seven tRNAs and all four rRNAs in the IR region. Of these genes, 16 contained one intron and two (*clpP* and *ycf3*) contained two introns.

The alignment of the 39 P. *chinensis* plastomes was 161,388 bp in length, and included 460 variable sites, 104 indels and three small inversions. The overall genetic diversity was 0.00082. Most of intraspecific *P. chinensis* variable sites and indels were located in the LSC and SSC regions (Figure 1), indicating that the IR region was more conserved than the single copy regions. The average number of intraspecific variable sites was 2.9 per kb and the and indel density was 0.65 per kb. Nucleotide diversity averaged over 500 bp windows showed that two intergenic regions of *trnH-psbA* and *ndhF-rpl32* had the highest sequence divergence (Figure 1).

Of the 104 indels in the 39 P. *chinensis* plastomes,72 were SSR-related indels, 21 were repeat-related indels, and 11 were normal indels. All SSR-related indels were located in the non-coding regions. The indel size ranged from 1 to 9 bp, and 1 bp indels was present at the highest frequency (58.3%). Except for

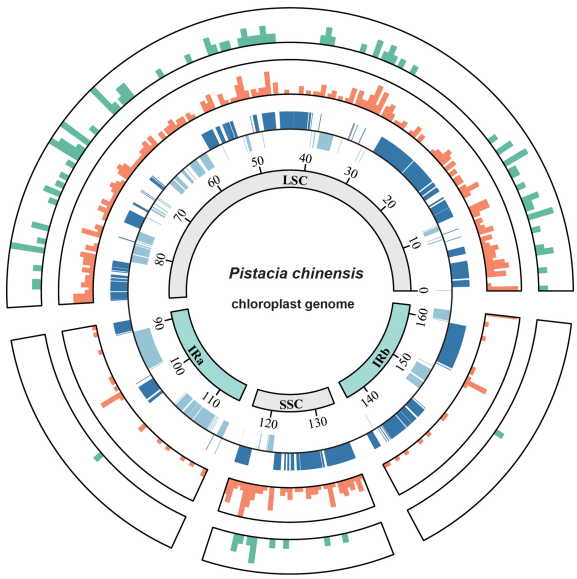


FIGURE 1
Circos plot showing the indel and nucleotide diversity of the *Pistacia chinensis* plastome. The concentric circles (inner to outer) indicate the following: quadripartite structure (represented by different colors); location of genes in the plastome; nucleotide diversity; and the number of indels. Nucleotide diversity and number of indels were computed for windows of 500 bp.

the indel in *clpP-psbB*, which was 3 bp long, all the normal indels were 1 bp long. The repeat-related indels ranged from 4 to 75 bp, with the longest occurring in the *ycf3-trnS* region, and found in an insert in the Yunnan01 sample.

All three small inversions formed stem-loop structures, and the lengths of these inversions were 3, 2, and 4 bp with the franking repeats of size 14, 22, and 14 bp, respectively. The lengths of the inversions and the flanking repeats were not correlated, which is consistent with previous research. The three small inversions were located in *atpF-atpH*, *petD-rpoA*, and *rpl14-rpl16*, and all of them occurred in the non-coding regions of the LSC.

Genetic diversity and intraspecific differentiation based on the plastome dataset

We inferred phylogenetic trees using ML and BI methods, based on whole plastome sequences. All the 39 samples were clearly divided into five clades (Figure 2B). Population structure results from ADMIXTURE suggested that there were six clades with K=6 (Figure 2C, Figure S2). The PCA results revealed three major groups (Figure 2A).

The first principal component explained 40.43% of total variance and clearly separated Clade I and Clade II. In total, 12

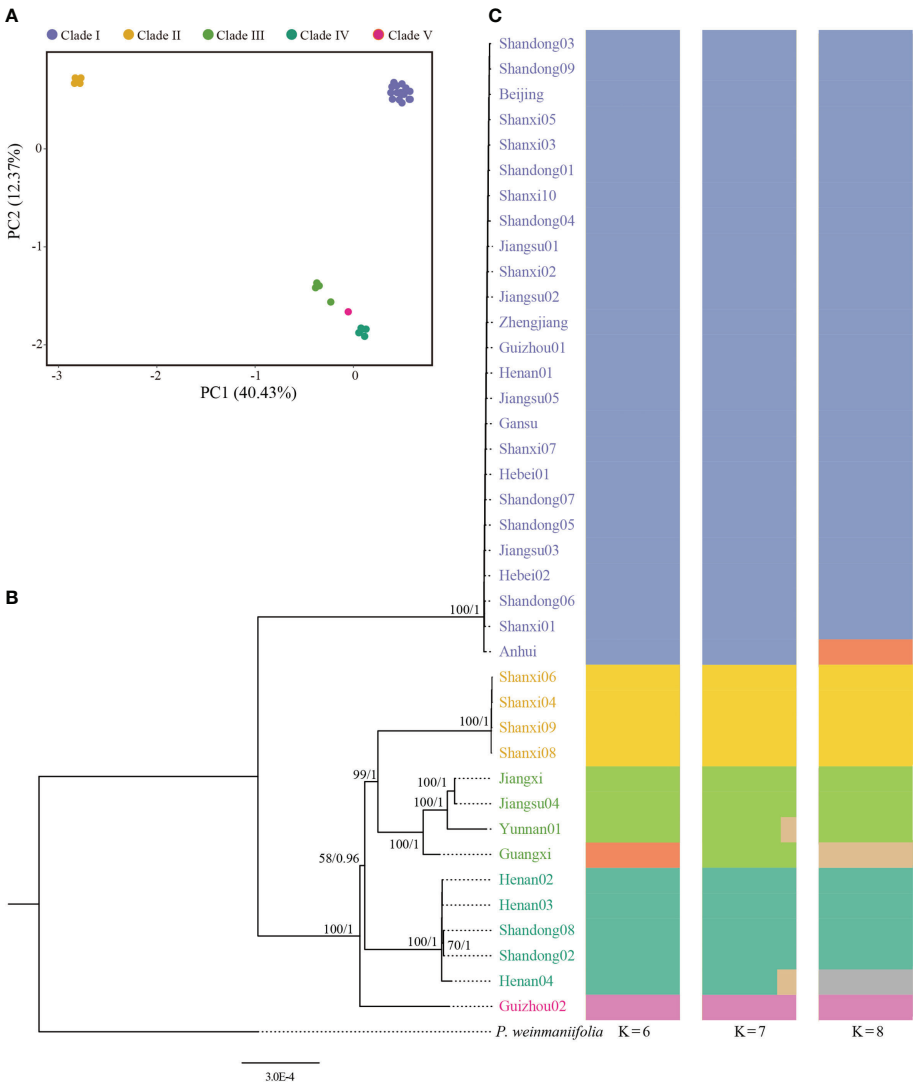


FIGURE 2
Genetic diversity of *Pistacia chinensis* assessed using complete plastome sequences. **(A)** Principal component analysis, **(B)** Phylogenetic tree. ML bootstrap support values/Bayesian posterior probabilities are shown at each node. **(C)** Population structure analysis with K = 6, 7 and 8.

distinct plastid haplotypes were identified, differing by between one and 279 plastid SNPs (Figure 3). The network of plastid haplotypes supported there were five clades.

The first clade contained 25 samples from Shandong, Beijing, Shanxi, Jiangsu, Zhejiang, Gansu, Guizhou, Hebei, and Anhui. This clade contained three haplotypes (Hap 1, Hap 2 and Hap 3), with Hap 1 being the most common. Clade I, which was sister to the other clades, exhibited significant genetic difference from the other clades and had the highest number of mutational steps (195). Clades II and III formed a single, highly supported (BS/PP=100/1) clade, and was sister to Clade IV which had lower supported (BS/PP=58/0.96) (Figure 2B). Clade II contained four samples from Shanxi with a single genotype (Hap 4). Clade III contained four samples from Jiangxi, Jiangsu, Yunnan, and Guangxi. The four samples showed significant divergence and included four different haplotypes (Hap 8-Hap 11). Clade IV contained five samples from Henan and Shandong. A single sample from Guizhou (Guizhou02, from Rongjiang County, Qiandongnan, Guizhou) formed Clade V.

Genetic diversity and intraspecific differentiation based on the nuclear SNPs dataset

Using the nuclear genome of *Pistacia vera* as the reference genome, we identified 3,632,308 SNPs with less than 10% missing data. The dataset of nuclear SNPs included only those sites that were polymorphic among the 39 sampled individuals. Population structure was analyzed using K values ranging from 1 to 10; the populations were clearly divided into two clades with K = 2, while the cross validation (CV) error was also the lowest with K = 2 (Figure 4C, Figure S2). The stu01 group contained nine samples from Jiangsu and Shandong. The stu02 group included 15 individuals from Gansu, Guangxi, Guizhou, Henan, Shanxi, and Yunnan. The remaining 15 individuals were classified as hybrids (the "cross" group), on the basis of the admixture coefficient according to the population structure at K = 2. The TreeMix results also identified strong gene flow from a node clustering stu01 and cross group into population cross (Figure 4A). The PCA results was showed in Figure S3.

Phylogenetic analysis of the 39 P. *chinensis* samples was performed based on the all the SNPs using the ML method (Figure 4B). Most of the nodes were well supported. Four samples (Guizou01, Guizhou02, Yunnan01, and Guangxi) formed a clade was the earliest diverged group. The two groups (stu01 and stu02) identified from ADMIXTURE did not form a monophyletic group.

Discordance relationships between nuclear and plastomes

In order to discover the discordance relationships between nuclear and plastome dataset, we compared the two phylogenetic trees (Figure 5). The 15 samples were deleted in

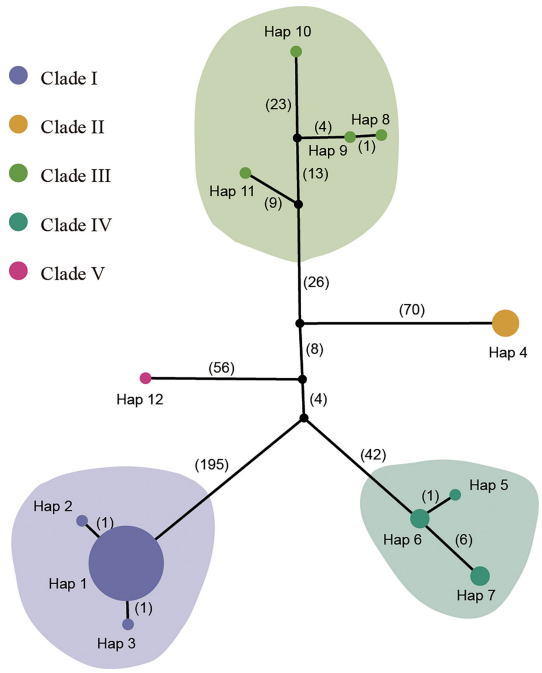


FIGURE 3
TCS network of 12 haplotypes from the plastome sequences. The colored circles represent plastid haplotypes; the black circles are extinct haplotypes; the number of mutational steps is shown on the lines. The size of the pie chart represents the number of the accessions. The haplotype for each sample is listed in Table S1.

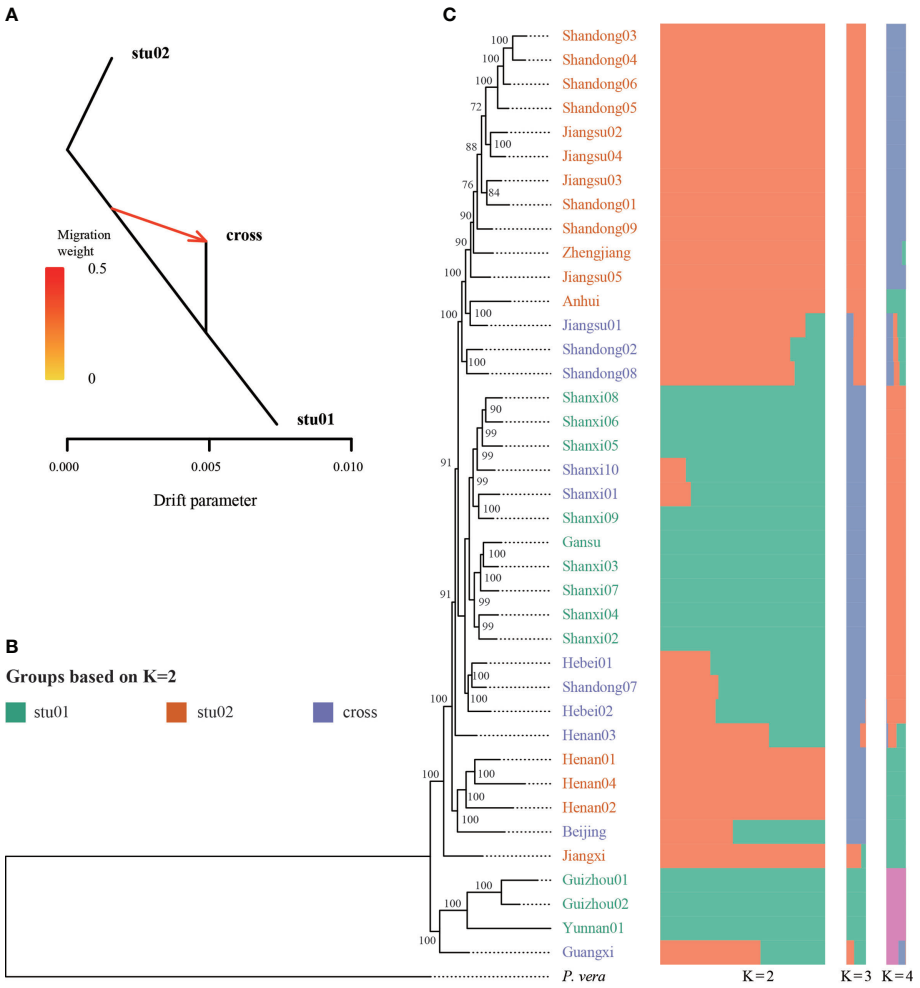


FIGURE 4
Genetic diversity of *Pistacia chinensis* based on nuclear SNPs. **(A)** Signal of introgression among different groups detected using the TreeMix program. According to the structure results (K = 2), we defined three groups, stu01, stu02 and the crossed (hybrid) samples. **(B)** Phylogenetic tree. ML bootstrap support values are shown at each node. **(C)** Population structure analysis with K = 2, 3 and 4.

this analysis which were classified as hybrids according to the results of the population structure at K = 2. For the *P. chinensis* samples, more closely related samples according to their plastomes tend to share identical or more similar haplotypes of their nuclear genomes (Figure 5), suggesting co-evolution between the plastomes and the nuclear genomes as a general pattern. However, more apparent exceptions were also observed. For example, the two samples of Shanxi07 and Guizhou01 with highly diverged nuclear genomes were detected to share identical haplotype of the plastomes, and the two samples of Anhui and Jiangxi with more closely related nuclear genomes were detected to have more diverged haplotypes of the plastomes. The discordance phylogeny relationships between nuclear and plastomes suggest that hybridization events between highly diverged samples within the *P. chinensis* subclades have also occurred, and such events are likely to be responsible for the observed discordance between the nuclear and plastomes.

Divergence time of *Pistacia chinensis*

Divergence time estimates showed that the stem and crown nodes of *Pistacia* were 37.74 Ma (95% highest posterior density (HPD): 35.73-39.87) in the later of Eocene and 15.68 Ma (95% HPD: 6.7-26.65) in the middle Miocene, respectively (Figure 6). Phylogenetic inference of plastome haplotypes subdivided 12 haplotypes into five main clades. Molecular dating analysis suggested that the firstly diverged during the later Miocene,

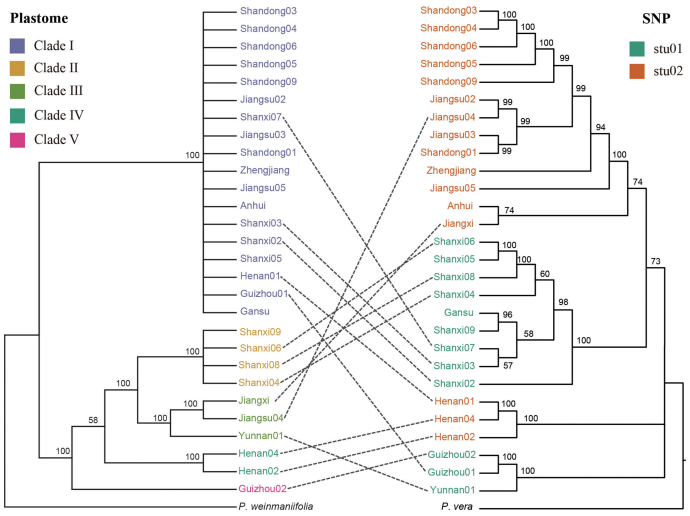


FIGURE 5
Discordance between phylogenetic trees reconstructed using plastomes (left) and nuclear SNPs (right). The trees included all the samples from the stu01 and stu02 clades from Figure 4B.

8.42 Ma (95% HPD: 3.48-14.97). The second divergence was occurred in the 4.82 Ma (95% HPD: 2.08-8.16) in the middle of Pliocene, giving the clade V. The crown age of Clade II, Clade III, and Clade IV and the divergence time between Clade II and Clade III were also in the middle of Pliocene. The divergence time of different genotypes within the clade was occurred in the Pleistocene.

Discussion

The plastome of *Pistacia chinensis* is highly variable

With the advent of next generation sequencing, plastome sequence data has become a basic tools and is extensively used to

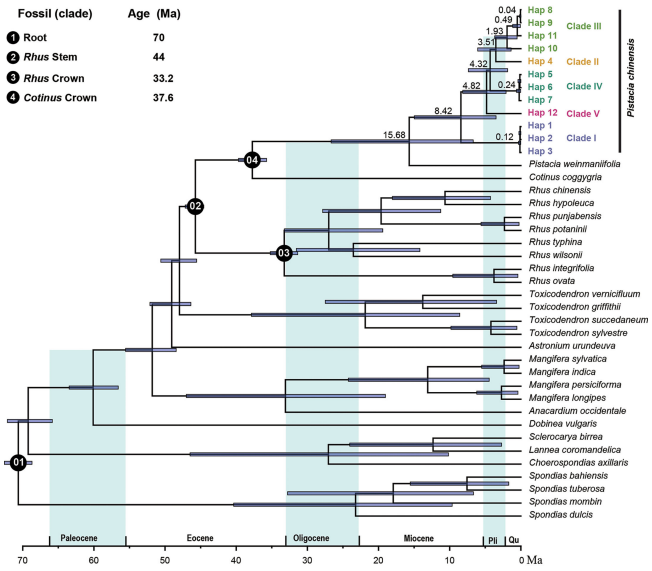


FIGURE 6
Divergence time of Anacardiaceae. The mean divergence time of the nodes is shown next to the nodes, and the blue bars correspond to the 95% highest posterior density (HPD).

resolve evolutionary relationships among plant species at different taxonomic levels (Sloan et al. 2014; Dong et al., 2018; Wikström et al., 2020; Zhao et al., 2021; Dong et al., 2021a; Dong et al., 2021c). Moreover, the capacity to assess the mutation rate or detect variation has significantly improved. Many studies have focused on variations at the species level or above (Liu et al., 2021; Ren et al., 2021; Tian et al., 2021; Xiong et al., 2021), however, only few studies have examined intraspecific diversity for whole plastome sequences (Hu et al., 2022; Lian et al., 2022; Sun et al., 2022). Plastid markers have been used in tree population genetics but were often limited to few polymorphic sites (Mariotti et al., 2010; Huang et al., 2014; Mohamoud et al., 2019). In this study, we sequenced the plastomes of 39 individuals of *Pistacia chinensis* sampled from the germplasm bank, and assessed the variation in the genomes. In total, 460 intraspecific polymorphic sites, 104 indels and three small inversions were identified among these 39 individuals. Mutation diversity of the plastome in *P. chinensis* is relatively high compared to that published for other plant species, including *Arnebia guttata* (313 polymorphic sites, 17 individuals) (Sun et al., 2022), *Bretschneidera sinensis* (105 polymorphic sites, 55 indels, 12 individuals) (Shang et al., 2022), the model grass plant *Brachypodium distachyon* (298 polymorphic sites, 53 individuals) (Sancho et al., 2018), and *Ginkgo biloba* (135 polymorphic sites, 71 individuals) (Hohmann et al., 2018).

Mutation rate variation among different lineages of plastomes has been examined in various studies (Smith and Donoghue, 2008; Schwarz et al., 2017; Choi et al., 2021). A hypothesis commonly invoked is that mutation rates are negatively correlated with generation time (Dong et al., 2022). For example, long-lived woody plants have lower mutation rates that do short-lived herbaceous species (Smith and Donoghue, 2008; Amanda et al., 2017; Dong et al., 2022), and the reed canary grass *Phalaris arundinacea* has high intraspecific plastid diversity (Perdereau et al., 2017). However, the relatively high intraspecific mutation rate we see in *Pistacia chinensis*, a long-lived plant, is not satisfactorily explained by the generation time hypothesis. Divergence time estimation indicated that *P. chinensis* speciated very early, during the late Miocene, 8.42 Ma (Figure 6). Relatively high genetic diversity is therefore likely to be a consequence of persistence of genetically distinct populations through periods of historical climate variability during the long evolutionary history. From the Pliocene to the Early Pleistocene, Eastern China was affected by the uplift of the northeastern and Southeastern Tibet Plateau (An et al., 2006), and the intensification of the East Asian summer monsoon (EASM) (An et al., 2014) and South Asian summer monsoon (SASM) (Chang et al., 2010) also occurred during this time.

Regions of the plastome with higher mutation rates (so-called mutational hotspot regions) have been observed in previous research, and the IR regions are known to be more conserved than the LSC and SSC regions (Dong et al., 2021b; Dong et al., 2021c). In the *P. chinensis* plastome, intraspecific variable sites and indels were mostly located in the LSC and SSC regions. Regions of particularly high variability in *P. chinensis* included *trnH-psbA* and *ndhF-rpl32*. Both have been identified as universal and variable markers suitable for intraspecific level studies, such as investigations into genetic diversity or population structure.

Genetic diversity in *Pistacia chinensis*

Using plastid and nuclear genome data, we provided much needed information on the genetic diversity of *P. chinensis*. Population structure results revealed five and two clusters of *P. chinensis* with high levels of diversity, using the plastid and nuclear genomes, respectively (Figures 2 and 4). High genetic differentiation was detected between distant clusters, however, these clusters did not reflect geography, suggesting that geographic distance did not explain the patterns in genetic structure according to the plastome dataset. For example, for the Clade I in the plastome dataset, this clade included the samples from nine provinces of China. The Clade III included four samples, which located in the four separate provinces of China (Guangxi, Jiangxi, Jiangsu, and Yunnan) (Figure 2B).The SSR results found similar results for the eight sampled *P. chinensis* populations, with different populations showing high differentiation and some of the distant populations showing high genetic similarity (Wu et al., 2010). Other factors may therefore be driving the population structure of *P. chinensis*, such as gene flow or introgression. When exclude the samples with hybrids, the samples in the stu01 group were most located in the east of China, and the stu02 group were most located in the west of China (Figure 5). This indicated both clusters reflected the geography excluding the effective of gene flow or introgression.

Most individuals at the tips of the phylogeny and population structure displayed apparent discordance between their nuclear and plastomes. This can be caused by factors such as hybridization and/or introgression between highly divergent populations (Wang et al., 2019; Cui et al., 2020). Most of the individuals in the *P. chinensis* subclade diverged between the late Miocene and the mid-Pliocene. The southern Chinese distribution of *P. chinensis* may have been subject to climate change and intensification of the EASM and SASM, creating opportunities for mixing and introgression between the early divergent clades.

On top of the evolutionary factors, cultivation may be a major factor effecting the genetic diversity of *P. chinensis*. Genetic diversity is generally thought to decrease with cultivation (Varshney et al., 2021; Wei and Jiang, 2021), as not all genotypes will be retained. Meanwhile, hybridization under artificial intervention or natural hybridization further leads to the mixing of wild resources. High heterozygosity is one characteristic of many cultivated plants and one of many

recognized challenges facing plant breeding. For *P. chinensis*, one third of the sampled individuals were identified as hybrids. Clade I from the plastid data included 25 samples from Shandong, Beijing, Shanxi, Jiangsu, Zhejiang, Gansu, Guizhou, Hebei, and Anhui. The plastome sequences of these samples were very similar to each other, suggesting that this clade may comprise the cultivated individuals. Our results indicated that during the cultivation process, the genetic diversity of *P. chinensis* may have decreased, suggesting that more genotypes, which are potentially of use in further breeding of this important species, should be conserved.

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: https://www.ncbi.nlm.nih.gov/genbank/, OP554519- OP554557.

Author contributions

BH, X-MX and W-QL conceived and designed the study. BH, M-JZ, YX, C-CC, HX and DL collected and analyzed the data. BH wrote the manuscript. LW edited and improved the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the Project funded by the Postdoctoral Science Foundation "Research and development of key technologies and equipment of germplasm bank" (BSHCX202101), the Postdoctoral Station Recruitment Subsidy of Shandong Province "Collection, preservation, evaluation and utilization of *Quercus acutissima* and *Q. variabilis* Germplasm Resources" (BSHCX202102).

Acknowledgments

We appreciate the facilitation provided by National Wild Plant Germplasm Resource Center. And we thank Dr. Jane Marczewski for polishing the English text professionally.

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.2022.1030647/full#supplementary-material

SUPPLEMENTARY TABLE 1
The haplotypes of 39 *Pistacia chinensis* samples.

SUPPLEMENTARY FIGURE 1
Gene maps of the *Pistacia chinensis* plastome. The genes are color-coded based on their functions. The dashed area indicates the GC composition of the plastome.

SUPPLEMENTARY FIGURE 2
The results of CV error for the plastome and SNPs datasets.

SUPPLEMENTARY FIGURE 3
Principal component analysis based on the nuclear SNPs dataset.

References

Alexander, D. H., and Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinf.* 12, 246. doi: 10.1186/1471-2105-12-246

Amanda, R., Li, Z., Van De Peer, Y., and Ingvarsson, P. K. (2017). Contrasting rates of molecular evolution and patterns of selection among gymnosperms and flowering plants. *Mol. Biol. Evol.* 34, 1363–1377. doi: 10.1093/molbev/msx069

An, Z., Sun, Y., Zhou, W., Liu, W., Qiang, X., Wang, X., et al (2014). "Chinese Loess and the East Asian Monsoon," In *Late Cenozoic Climate Change in Asia: Loess, Monsoon and Monsoon-arid Environment Evolution*, ed. Z. An. (Dordrecht: Springer Netherlands) 23-143.

An, Z., Zhang, P., Wang, E., Wang, S., Qiang, X., Li, L., et al. (2006). Changes of the monsoon-arid environment in China and growth of the Tibetan plateau since the Miocene. *Quaternary Sci.* 26, 678–693. doi: 1001-7410(2006)05-678-16

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

Bouckaert, R., Heled, J., Kuhnert, D., Vaughan, T., Wu, C. H., Xie, D., et al. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* 10, e1003537. doi: 10.1371/journal.pcbi.1003537

Chang, Z., Xiao, J., Lü, L., and Yao, H. (2010). Abrupt shifts in the Indian monsoon during the pliocene marked by high-resolution terrestrial records from

the yuanmou basin in southwest China. *J. Asian Earth Sci.* 37, 166–175. doi: 10.1016/j.jseaes.2009.08.005

Cheng, X., Wang, F., Luo, W., Kuang, J., and Huang, X. (2022). Transcriptome analysis and identification of a female-specific SSR marker in pistacia chinensis based on illumina paired-end RNA sequencing. *Genes* 13 (6), 1024. doi: 10.3390/genes13061024

Choi, K., Weng, M.-L., Ruhlman, T. A., and Jansen, R. K. (2021). Extensive variation in nucleotide substitution rate and gene/intron loss in mitochondrial genomes of pelargonium. *Mol. Phylogenet. Evol.* 155, 106986. doi: 10.1016/j.ympev.2020.106986

Clement, M., Snell, Q., Walker, P., Posada, D., and Crandall, K. (2002). TCS: Estimating gene genealogies. *Parallel Distributed Process. Symposium Int. Proc.* 2, 184. doi: 10.1109/IPDPS.2002.1016585

Cui, H., Ding, Z., Zhu, Q., Wu, Y., and Gao, P. (2020). Population structure and genetic diversity of watermelon (Citrullus lanatus) based on SNP of chloroplast genome. *3 Biotech.* 10, 374. doi: 10.1007/s13205-020-02372-5

Dong, W., Li, E., Liu, Y., Xu, C., Wang, Y., Liu, K., et al. (2022). Phylogenomic approaches untangle early divergences and complex diversifications of the olive plant family. *BMC Biol.* 20, 92. doi: 10.1186/s12915-022-01297-0

Dong, W., Liu, Y., Xu, C., Gao, Y., Yuan, Q., Suo, Z., et al. (2021a). Chloroplast phylogenomic insights into the evolution of distylium (Hamamelidaceae). *BMC Genomics* 22, 293. doi: 10.1186/s12864-021-07590-6

Dong, W., Sun, J., Liu, Y., Xu, C., Wang, Y., Suo, Z., et al. (2021b). Phylogenomic relationships and species identification of the olive genus Olea (Oleaceae). *J. Systematics Evol.* doi: 10.1111/jse.12802

Dong, W., Xu, C., Liu, Y., Shi, J., Li, W., and Suo, Z. (2021c). Chloroplast phylogenomics and divergence times of lagerstroemia (Lythraceae). *BMC Genomics* 22, 434. doi: 10.1186/s12864-021-07769-x

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

Fu, R., Zhu, Y., Liu, Y., Feng, Y., Lu, R.-S., Li, Y., et al. (2022). Genome-wide analyses of introgression between two sympatric Asian oak species. *Nat. Ecol. Evol.* 6, 924–935. doi: 10.1038/s41559-022-01754-7

Heldenbrand, J. R., Baheti, S., Bockol, M. A., Drucker, T. M., Hart, S. N., Hudson, M. E., et al. (2019). Recommendations for performance optimizations when using GATK3.8 and GATK4. *BMC Bioinf.* 20, 557. doi: 10.1186/s12859-019-3169-7

Hohmann, N., Wolf, E. M., Rigault, P., Zhou, W., Kiefer, M., Zhao, Y., et al. (2018). Ginkgo biloba's footprint of dynamic pleistocene history dates back only 390,000 years ago. *BMC Genomics* 19, 299. doi: 10.1186/s12864-018-4673-2

Huang, D. I., and Cronk, Q. C. B. (2015). Plann: A command-line application for annotating plastome sequences. *Appl. Plant Sci.* 3, 1500026. doi: 10.3732/apps.1500026

Huang, D. I., Hefer, C. A., Kolosova, N., Douglas, C. J., and Cronk, Q. C. (2014). Whole plastome sequencing reveals deep plastid divergence and cytonuclear discordance between closely related balsam poplars, populus balsamifera and p. trichocarpa (Salicaceae). *New Phytol.* 204, 693–703. doi: 10.1111/nph.12956

Hu, G., Wu, Y., Guo, C., Lu, D., Dong, N., Chen, B., et al. (2022). Haplotype analysis of chloroplast genomes for jujube breeding. *Front. Plant Sci.* 13, 841767. doi: 10.3389/fpls.2022.841767

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

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

Karbstein, K., Tomasello, S., Hodač, L., Wagner, N., Marinček, P., Barke, B. H., et al. (2022). Untying Gordian knots: unraveling reticulate polyploid plant evolution by genomic data using the large ranunculus auricomus species complex. *New Phytol.* 235, 2081-2098. doi: 10.1111/nph.18284

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

Katsiotis, A., Hagidimitriou, M., Drossou, A., Pontikis, C., and Loukas, M. (2003). Genetic relationships among species and cultivars of pistacia using RAPDs and AFLPs. *Euphytica* 132, 279–286. doi: 10.1023/A:1025027323184

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35, 4453–4455. doi: 10.1093/bioinformatics/btz305

Kumar, P., Gupta, V. K., Misra, A. K., Modi, D. R., and Pandey, B. K. (2009). Potential of molecular markers in plant biotechnology. *Plant Omics* 2, 141–162.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Leigh, J. W., and Bryant, D. (2015). POPART: full-feature software for haplotype network construction. *Methods Ecol. Evol.* 6, 1110–1116. doi: 10.1111/2041-210X.12410

Lian, C., Yang, H., Lan, J., Zhang, X., Zhang, F., Yang, J., et al. (2022). Comparative analysis of chloroplast genomes reveals phylogenetic relationships and intraspecific variation in the medicinal plant isodon rubescens. *PLoS One* 17, e0266546. doi: 10.1371/journal.pone.0266546

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics* 26, 589–595. doi: 10.1093/bioinformatics/btp698

Li, X., Hou, S., Su, M., Yang, M., Shen, S., Jiang, G., et al. (2010). Major energy plants and their potential for bioenergy development in China. *Environ. Manage.* 46, 579–589. doi: 10.1007/s00267-010-9443-0

Liu, S., Wang, Z., Su, Y., and Wang, T. (2021). Comparative genomic analysis of polypodiaceae chloroplasts reveals fine structural features and dynamic insertion sequences. *BMC Plant Biol.* 21, 31. doi: 10.1186/s12870-020-02800-x

Li, J., Wang, S., Jing, Y., Wang, L., and Zhou, S. (2013). A modified CTAB protocol for plant DNA extraction. *Chin. Bull. Bot.* 48, 72–78. doi: 10.3724/SP.J.1259.2013.00072

Lu, J.-T., Qiu, Y.-H., and Lu, J.-B. (2019). Effects of landscape fragmentation on genetic diversity of Male-biased dioecious plant pistacia chinensis bunge populations. *Forests* 10:792. doi: 10.3390/f10090792

Magdy, M., Ou, L., Yu, H., Chen, R., Zhou, Y., Hassan, H., et al. (2019). Pan-plastome approach empowers the assessment of genetic variation in cultivated capsicum species. *Horticulture Res.* 6, 108. doi: 10.1038/s41438-019-0191-x

Manchester, S. R. (1994). Fruits and seeds of the middle Eocene nut beds flora, clarno formation, Oregon. *Paleoutographica Americana* 58, 1–205.

Mariotti, R., Cultrera, N. G. M., Díez, C. M., Baldoni, L., and Rubini, A. (2010). Identification of new polymorphic regions and differentiation of cultivated olives (Olea europaea l.) through plastome sequence comparison. *BMC Plant Biol.* 10, 211. doi: 10.1186/1471-2229-10-211

Migliore, J., Kaymak, E., Mariac, C., Couvreur, T. L. P., Lissambou, B. J., Piñeiro, R., et al. (2019). Pre-pleistocene origin of phylogeographical breaks in African rain forest trees: New insights from greenway odendron (Annonaceae) phylogenomics. *J. Biogeography* 46, 212–223. doi: 10.1111/jbi.13476

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., et al. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37, 1530–1534. doi: 10.1093/molbev/msaa015

Mohamoud, Y. A., Mathew, L. S., Torres, M. F., Younuskunju, S., Krueger, R., Suhre, K., et al. (2019). Novel subpopulations in date palm (Phoenix dactylifera) identified by population-wide organellar genome sequencing. *BMC Genomics* 20, 498. doi: 10.1186/s12864-019-5834-7

Parfitt, D. E., and Badenes, M. L. (1997). Phylogeny of the genus pistacia as determined from analysis of the chloroplast genome. *Proc. Natl. Acad. Sci.* 94, 7987–7992. doi: 10.1073/pnas.94.15.7987

Perdereau, A., Klaas, M., Barth, S., and Hodkinson, T. R. (2017). Plastid genome sequencing reveals biogeographical structure and extensive population genetic variation in wild populations of phalaris arundinacea l. @ in north-western Europe. *GCB Bioenergy* 9, 46–56. doi: 10.1111/gcbb.12362

Pickrell, J. K., and Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *Nat. Precedings* 48. doi: 10.1038/npre.2012.6956.1

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender, D., et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Rambaut, A. (1996). Se-Al: sequence alignment editor. version 2.0. Available at: http://tree.bio.ed.ac.uk/software/seal/.

Rambaut, A., Suchard, M., Xie, D., and Drummond, A. (2014). *Tracer v1. 6*. Available at: http://beast.bio.ed.ac.uk

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

Rozas, J., Ferrer-Mata, A., Sanchez-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

Sancho, R., Cantalapiedra, C. P., Lopez-Alvarez, D., Gordon, S. P., Vogel, J. P., Catalan, P., et al. (2018). Comparative plastome genomics and phylogenomics of brachypodium: flowering time signatures, introgression and recombination in recently diverged ecotypes. *New Phytol.* 218, 1631–1644. doi: 10.1111/nph.14926

Schwarz, E. N., Ruhlman, T. A., Weng, M.-L., Khiyami, M. A., Sabir, J. S. M., Hajarah, N. H., et al. (2017). Plastome-wide nucleotide substitution rates reveal

accelerated rates in papilionoideae and correlations with genome features across legume subfamilies. *J. Mol. Evol.* 84, 187–203. doi: 10.1007/s00239-017-9792-x

Shang, C., Li, E., Yu, Z., Lian, M., Chen, Z., Liu, K., et al. (2022). Chloroplast genomic resources and genetic divergence of endangered species bretschneidera sinensis (Bretschneideraceae). *Front. Ecol. Evol.* 10. doi: 10.3389/fevo.2022.873100

Sloan, D. B., Triant, D. A., Forrester, N. J., Bergner, L. M., Wu, M., and Taylor, D. R. (2014). A recurring syndrome of accelerated plastid genome evolution in the angiosperm tribe Sileneae (Caryophyllaceae). *Mol. Phylogenet. Evol* 72, 82-89. doi: 10.1016/j.ympev.2013.12.004

Smith, S. A., and Donoghue, M. J. (2008). Rates of molecular evolution are linked to life history in flowering plants. *Science* 322, 86–89. doi: 10.1126/science.1163197

Sun, J., Wang, S., Wang, Y., Wang, R., Liu, K., Li, E., et al. (2022). Phylogenomics and genetic diversity of arnebiae radix and its allies (Arnebia, boraginaceae) in China. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.920826

Tang, M., Zhang, P., Zhang, L., Li, M., and Wu, L. (2012). A potential bioenergy tree: Pistacia chinensis bunge. *Energy Proc.* 16, 737–746. doi: 10.1016/j.egypro.2012.01.119

Tian, X., Guo, J., Zhou, X., Ma, K., Ma, Y., Shi, T., et al. (2021). Comparative and evolutionary analyses on the complete plastomes of five kalanchoe horticultural plants. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.705874

Varshney, R. K., Roorkiwal, M., Sun, S., Bajaj, P., Chitikineni, A., Thudi, M., et al. (2021). A chickpea genetic variation map based on the sequencing of 3,366 genomes. *Nature* 599, 622–627. doi: 10.1038/s41586-021-04066-1

Wang, X., Chen, L., and Ma, J. (2019). Genomic introgression through interspecific hybridization counteracts genetic bottleneck during soybean domestication. *Genome Biol.* 20, 22. doi: 10.1186/s13059-019-1631-5

Wang, L., Yu, H., and He, X. (2012). Assessment on fuel properties of four woody biodiesel plants species in China. *Scientia Silvae Sinicae* 48, 150–154.

Wei, X., and Jiang, M. (2021). Meta-analysis of genetic representativeness of plant populations under ex situ conservation in contrast to wild source populations. *Conserv. Biol.* 35, 12–23. doi: 10.1111/cobi.13617

Wikström, N., Bremer, B., and Rydin, C. (2020). Conflicting phylogenetic signals in genomic data of the coffee family (Rubiaceae). *J. Syst. Evol* 58, 440-460. doi: 10.1111/jse.12566

Wu, Z., Zhang, Z., Wang, Z., Li, J., and Wang, X. (2010). SSR analysis on genetic diversity of natural populations of pistacia chinensis bunge. *Chin. J. Appl. Environ. Biol.* 16, 803–806. doi: 10.3724/SP.J.1145.2010.00803

Xie, L., Yang, Z.-Y., Wen, J., Li, D.-Z., and Yi, T.-S. (2014). Biogeographic history of pistacia (Anacardiaceae), emphasizing the evolution of the madrean-tethyan and the eastern Asian-tethyan disjunctions. *Mol. Phylogenet. Evol.* 77, 136–146. doi: 10.1016/j.ympev.2014.04.006

Xiong, Q., Hu, Y., Lv, W., Wang, Q., Liu, G., and Hu, Z. (2021). Chloroplast genomes of five oedogonium species: genome structure, phylogenetic analysis and adaptive evolution. *BMC Genomics* 22, 707. doi: 10.1186/s12864-021-08006-1

Xue, C., Geng, F. D., Li, J. J., Zhang, D. Q., Gao, F., Huang, L., et al. (2021). Divergence in the aquilegia ecalcarata complex is correlated with geography and climate oscillations: Evidence from plastid genome data. *Mol. Ecol.* 30, 5796–5813. doi: 10.1111/mec.16151

Zeng, L., Tu, X.-L., Dai, H., Han, F.-M., Lu, B.-S., Wang, M.-S., et al. (2019). Whole genomes and transcriptomes reveal adaptation and domestication of pistachio. *Genome Biol.* 20, 79. doi: 10.1186/s13059-019-1686-3

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

Zheng, S., Poczai, P., Hyvonen, J., Tang, J., and Amiryousefi, A. (2020). Chloroplot: An online program for the versatile plotting of organelle genomes. *Front. Genet.* 11, 576124. doi: 10.3389/fgene.2020.576124

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 14 November 2022
DOI 10.3389/fpls.2022.1059379

OPEN ACCESS

EDITED BY
Yu Song,
University of Chinese Academy of Sciences, China

REVIEWED BY
Xin Yao,
Xishuangbanna Tropical Botanical Garden (CAS), China
Chao Liu,
Qujing Normal University, China

*CORRESPONDENCE
Jin Cheng
chengjin@bjfu.edu.cn
Lei Xie
xielei@bjfu.edu.cn

†These authors have contributed equally to this work

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

RECEIVED 01 October 2022
ACCEPTED 24 October 2022
PUBLISHED 14 November 2022

CITATION
Xiao JM, Lyu RD, He J, Li MY, Ji JX, Cheng J and Xie L (2022) Genome-partitioning strategy, plastid and nuclear phylogenomic discordance, and its evolutionary implications of *Clematis* (Ranunculaceae). *Front. Plant Sci.* 13:1059379. doi: 10.3389/fpls.2022.1059379

Genome-partitioning strategy, plastid and nuclear phylogenomic discordance, and its evolutionary implications of *Clematis* (Ranunculaceae)

Jiamin Xiao[1†], Rudan Lyu[1†], Jian He[1], Mingyang Li[2], Jiaxin Ji[1], Jin Cheng[2*] and Lei Xie[1*]

[1]School of Ecology and Nature Conservation, Beijing Forestry University, Beijing, China, [2]College of Biological Sciences and Technology, Beijing Forestry University, Beijing, China

Clematis is one of the largest genera of Ranunculaceae with many phylogenetic problems left to be resolved. *Clematis* species have considerable genome size of more than 7 Gbp, and there was no whole-genome reference sequence published in this genus. This raises difficulties in acquiring nuclear genome data for its phylogenetic analysis. Previous studies based on Sanger sequencing data, plastid genome data, and nrDNA sequences did not well resolve the phylogeny of *Clematis*. In this study, we used genome skimming and transcriptome data to assemble the plastid genome sequences, nuclear single nucleotide polymorphisms (SNPs) datasets, and single-copy nuclear orthologous genes (SCOGs) to reconstruct the phylogenetic backbone of *Clematis*, and test effectiveness of these genome partitioning methods. We also further analyzed the discordance among nuclear gene trees and between plastid and nuclear phylogenies. The results showed that the SCOGs datasets, assembled from transcriptome method, well resolved the phylogenetic backbone of *Clematis*. The nuclear SNPs datasets from genome skimming method can also produce similar results with the SCOGs data. In contrast to the plastid phylogeny, the phylogeny resolved by nuclear genome data is more robust and better corresponds to morphological characters. Our results suggested that rapid species radiation may have generated high level of incomplete lineage sorting, which was the major cause of nuclear gene discordance. Our simulation also showed that there may have been frequent interspecific hybridization events, which led to some of the cyto-nuclear discordances in *Clematis*. This study not only provides the first robust phylogenetic backbone of *Clematis* based on nuclear genome data, but also provides suggestions of genome partitioning strategies for the phylogenomic study of other plant taxa.

KEYWORDS

***Clematis*, cyto-nuclear discordance, genome partitioning, genome skimming, transcriptome, phylogenomics**

Introduction

With the rapid development of molecular biotechnology, the cost of high-throughput sequencing continues to decrease. Using genomic data to reconstruct phylogeny and explore the origin and evolutionary history of plant taxa is growing rapidly (Zimmer and Wen, 2015; Wen et al., 2017; Marks et al., 2021; Kress et al., 2022). Compared to previous studies using the Sanger sequencing method, the application of genomic data has greatly improved the resolution of the phylogenetic trees (Valcárcel and Wen, 2019; Li et al., 2021; Khan et al., 2021). Genomic data can not only generate better resolved phylogenies of plant taxa, but can also alleviate the problem of stochastic error due to insufficient information from small datasets (Yu et al., 2018; Watson et al., 2020).

In recent years, the plastid genome (plastome) has been considered to be the most important source of data and widely applied for phylogenetic reconstruction of green plant phylogeny at almost all taxonomic levels (Li et al., 2019; Zhai et al., 2019; Zhang et al., 2020; Zhao et al., 2021). However, the uni-parental inherited plastid genome sometimes showed conflicting phylogenetic signals with the bi-parental inherited nuclear genome data (cyto-nuclear discordance) due to chloroplast capture, incomplete lineage sorting (ILS) (Rose et al., 2021), or other factors such as sampling error, stochastic error, paralogs, and so on (Zou and Ge, 2008). Comprehensive understanding of evolutionary process for a plant taxon requires both cytoplasmic and nuclear genome evidence and an in depth analysis of their phylogenetic discordance (Lee-Yaw et al., 2019).

Clematis L. is one of the largest genera in the family Ranunculaceae with about 300 wild species, most of which are diploid (Tamura, 1995; Wang and Bartholomew, 2001; Wang and Li, 2005). The taxonomy of *Clematis* has been considered to be difficult. Many classifications published in recent years held different views on many issues, including the delineation of the genus, infrageneric classification, and species delimitation (Tamura, 1995; Johnson, 1997; Grey-Wilson, 2000; Wang and Li, 2005). Previous molecular phylogenetic studies, based on the nuclear ribosomal DNA (nrDNA), the plastid fragments, and the complete plastome data, have solved many of those problems, such as genus delineation and the identification of the sister group of *Clematis* (Miikeda et al., 2006; Xie et al., 2011; Lehtonen et al., 2016; Jiang et al., 2017; He et al., 2021). However, all the previously published studies had limitations of not establishing a robust phylogenetic framework within *Clematis*, and its extensive cyto-nuclear discordance remains to be analyzed by inclusion of more nuclear genome data.

There are several reasons that may contribute to the difficulties in reconstructing a robust phylogeny of *Clematis*. Firstly, according to previous molecular studies, species radiation events may have happened during the late Neogene and the Quaternary (Xie et al., 2011; He et al., 2021). Small number of DNA sequences with insufficient informative loci often failed to resolve the relationships among recently radiated groups (Zhao et al., 2021). Secondly, interspecific hybridization may have happened or may be not uncommon in *Clematis* (Lyu et al., 2021), that may cause cyto-nuclear discordance during phylogenetic reconstruction. Thirdly, *Clematis* species have relatively large genome size (7.18 Gbp–16.43 Gbp, https://cvalues.science.kew.org/search) and there is no high-quality whole genome data available, which raise technical difficulties for genome-partitioning selection.

The genome-partitioning methods for phylogenomic study of plant taxa generally include reduced-representation Genome Sequencing (RRGS), genome skimming, transcriptome sequencing or RNA-seq, and target enrichment sequencing (Zimmer and Wen, 2015; Yu et al., 2018; McKain et al., 2018). Among them, genome skimming, which randomly captures a certain percentage of total genomic DNA (Dodsworth, 2015; Thode et al., 2020; Wikström et al., 2020), has been widely applied for phylogenetic studies (Wen et al., 2018; Su et al., 2021; He et al., 2021). One of the advantages of genome skimming method is that fresh, silica-gel dried, or even herbarium materials can be used for this method (Liu et al., 2019; Wang et al., 2020). Using genome skimming data, cytoplasmic genome and tandemly repeated nrDNA can be assembled for phylogenetic reconstruction (Yu et al., 2018; Fonseca and Lohmann, 2020). According to the recently developed method by Liu et al. (2021), genome skimming data with high sequencing depth (10 × or more) can be used for assembling single-copy nuclear genes for phylogenetic studies. Other studies have shown that genome skimming data with low sequencing depth (less than 1 ×) can be used to obtain single nucleotide polymorphisms (SNPs) from nuclear genome for phylogenetic reconstruction (Olofsson et al., 2019).

In contrast, transcriptome method has irreplaceable advantages for obtaining single-copy nuclear genes (One Thousand Plant Transcriptomes Initiative, 2019), and plant genome size is not the factor affecting sequencing depth because the transcribed gene content is small and very stable among seed plants (around 0.03 Gbp, Novák et al., 2020). However, the application of the transcriptome method is limited by plant material, which requires fresh plant tissue (or stored in RNA stabilization solution), or at least silica gel dried material (He et al., 2022). For a large genus like *Clematis*, a considerable proportion of species samples may be from herbarium specimens. It is difficult to obtain transcriptome data from all samples. In recent years, using targeted enrichment sequencing method to obtain nuclear gene has attracted much attention in phylogenetic studies (Vargas et al., 2019; Stull et al., 2020). This method can also use herbarium material for DNA extraction. However, comparing to RNA-seq method, the target enrichment method has much more complicated experimental process, relatively smaller amount of data, too much missing data, and low data reusability (McKain et al., 2018).

For *Clematis*, an accurate and well-supported phylogenetic backbone still remains to be reconstructed by nuclear genome data. Obtaining high-depth sequencing data (10 × means at least 70 Gbp for each sample in *Clematis*) to assemble nuclear genes is not economically viable for *Clematis*. In this study, using genome skimming (with low depth) and transcriptome data, we try to answer the following questions: to what extent nuclear genome data may improve the phylogenetic inference of *Clematis*? can genome skimming data with low sequencing depth provide more nuclear phylogenetic information? if the nuclear single nucleotide polymorphisms (SNPs) data from genome skimming method can be used for *Clematis* phylogenetic reconstruction? which one, incomplete lineage sorting or hybridization, may have caused the cyto-nuclear discordance of *Clematis*? This study will also shed light on the genome partitioning selection for phylogenomic analysis of other similar taxa with recent species radiation and considerable genome size.

Materials and methods

Plant material

Because the major purpose of this study is to check the robustness of the phylogenetic backbones inferred by different datasets, we chose a phylogenetically representative sampling scheme with only key species of *Clematis* in this study. A total of 32 species (about 1/10 of total species) were used for our phylogenomic analysis, covering all the subgenera of both Tamura (1995) and Wang and Li (2005). This sampling scheme also covers 11 sections (of the total 17) in the classification of Tamura (1995), and 9 sections (of the total 15) in the classification of Wang and Li (2005). Although we did not include several small sections (like sect. *Archiclematis*, sect. *Pterocarpa*, and sect. *Angustifoliae*), our sampling represented all the major lineages (clades) of *Clematis* included in previous studies (Miikeda et al., 2006; Xie et al., 2011; He et al., 2021). Furthermore, our previous studies showed that some sections, such as sect. *Clematis* and sect. *Viorna* (Reichb.) Prantl (sensu Wang and Li, 2005), may be polyphyletic. So, our samples also included problematic species of those sections (Supplementary Table S1).

The plant materials are mostly collected from the field, only with two samples from herbarium specimens. Among all the 32 sampled species, genome skimming data of 28 were newly generated for this study, and those of the other four species were retrieved from previous studies (Supplementary Table S1). Because specimen materials cannot yield RNA-seq data, transcriptomes of only 28 species were sequenced in this study. According to Jiang et al. (2017), *Anemoclema glaucifolium* (Franch.) W. T. Wang was chosen as an outgroup.

Methods for genomic data acquisition

Transcriptome sequencing

Transcriptome sequencing followed the method of He et al. (2022). Total RNAs were extracted at Biomarker Technologies Corporation (https://www.biomarker.com.cn) from silica gel dried leaves using TRIzon Reagent (TRIzon, CoWin Biosciences, Jiangsu, PR China). Then the RNAs were reversed into cDNA, and paired-end reads of 2 ×150 libraries were generated and sequenced on a NovaSeq 6000 platform (Illumina, San Diego, California, USA). About 6 Gbp of raw reads were obtained for each samples. The raw reads were then filtered and trimmed using fastp v.0.20 (Chen et al., 2018). The clean transcriptomes were *de novo* assembled using Trinity v.2.5.1 (Grabherr et al., 2011) with default parameters. All the transcriptome data were deposited in GenBank (Supplementary Table S2).

Genome skimming sequencing

The total genomic DNAs were extracted from silica-dried samples at Biomarker Technologies Corporation (https://www.biomarker.com.cn) using a genomic DNA extraction kit following manufacturer instructions (Tiangen Biotech Co. Ltd., Beijing, China). For the specimen samples, the total DNAs were obtained from the Herbarium of Institute of Botany, the Chinese Academy of Sciences (PE), and the extraction method was according to Li, 2013. Then, 2 ×150 bp paired-end libraries were constructed and sequenced using an illumina NovaSeq 6000 platform (Illumina, San Diego, California, USA). The newly sequenced samples yielded around 6 Gbp of raw data. In order to assemble the draft genome of *Clematis*, we extracted total DNA from a *C. brevicaudata* DC. sample and constructed a library for sequencing, finally obtaining raw data of about 200 Gbp. All the genome skimming data were deposited in GenBank (Supplementary Table S1).

Raw data processing

Plastid genome assembly

We used genome skimming data to assemble the complete plastid genome sequence using GetOrganelle v.1.7.5 (Jin et al., 2020). Detailed assembling process followed He et al. (2021). The assembled plastome sequences were annotated using Plann v.1.1.2 (Huang and Cronk, 2015) and manually adjusted by Geneious Prime v.2020 (Kearse et al., 2012).

Nuclear single-copy orthologous genes assembly using transcriptome data

Nuclear single-copy orthologs (SCOGs) were obtained from transcriptome data followed the pipeline of He et al. (2022). We used CD-HIT v.4.6.2 (Fu et al., 2012) to remove redundant sequences and TransDecoder v.5.0 (https://github.com/TransDecoder/TransDecoder/releases) to predict protein-coding regions. The assembly completeness of each sample was assessed using BUSCO v.5.2.2 (Simao et al., 2015). Subsequently, we constructed transcriptome homology scans using Proteinortho v.6.0.10 (Lechner et al., 2011) in the Diamond mode (Buchfink et al., 2015), and then searched the resulting clusters to identify gene families using a Python script "get_seq_from_proteinortho.py" (https://github.com/HeJian151004/get_seq_from_proteinortho). We then deleted all the organelle genome sequences from the SCOGs using the script "del_chloro_mito_from_fasta.py" (https://github.com/HeJian151004/del_chloro_mito_from_fasta), and used Treeshrink v.1.3.9 (Mai and Mirarab, 2018) to delete sequences that may be incorrectly clustered (showing unexpectedly long branches in the gene tree). Finally, we selected two SCOGs datasets with alignment length at least 1,000 bp (SCOG1000) and 3,000 bp (SCOG3000) for phylogenetic analysis.

Acquiring the nuclear SNPs data from genome skimming method

For the genome skimming data, we further mined the nuclear SNPs data for phylogenetic inference. In brief, we assembled a draft genome as a reference, and then mapped the genome skimming data of other species to this reference genome to obtain the SNPs dataset. We used two methods to obtain the SNPs data, the GATK and the Geneious pipelines. Detailed process of both pipelines are as follows (also shown in Figure 1).

First, we used the GATB-Minia (https://github.com/GATB/gatb-minia-pipeline) to assemble the draft genome (Drezen et al., 2014). We obtained a draft genome of 7.81Gbp, which is too large to be applied for downstream analysis. Therefore, we used the RepeatMasker v.4.0.9 (Chen, 2004) to exclude the repetitive regions in the draft genome. We further deleted the low coverage regions by the following processes: the genome skimming data of five distantly related *Clematis* species [*C. leschenaultiana* DC., *C. repens* Finet et Gagnep., *C. songorica* Bunge, *C. tibetana* Kuntze, *C. viridis* (W. T. Wang and M. C.

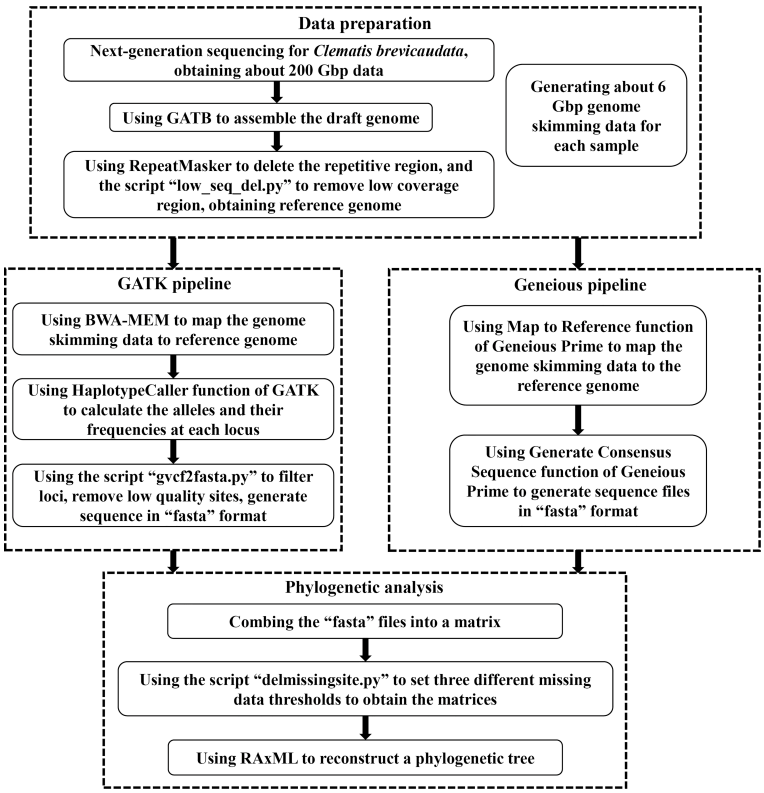


FIGURE 1
A flow chart of acquiring nuclear single nucleotide polymorphisms (SNPs) dataset from genome skimming data in this study.

Chang) W. T. Wang] were mapped to the draft genome by Map to Reference function of Geneious Prime v.2020 (Kearse et al., 2012). Then we used a script "low_seq_del.py" (https://github.com/Jhe1004/low_seq_del) to remove the regions that none of the five samples were matched. After removing the duplicate and low coverage regions, we finally obtained a reference genome of 616 Mbp.

The GATK pipeline used BWA-MEM v.0.7.1 (Li, 2013) to map each genome skimming data back to the reference genome to generate "bam" format files. Then the HaplotypeCaller function of GATK v.4.2.5 (Mckenna et al., 2010) was applied to calculate the alleles and their frequencies at each locus. Then, GATK output the result as the "vcf" format file. Then, we used script "gvcf2fasta.py" (https://github.com/Jhe1004/gvcf2fasta) to convert "vcf" file to the "fasta" sequence. We filtered and deleted the site that met any of the following three criteria: (1) coverage less than 4, (2) site quality score less than 20, and (3) heterozygous.

The Geneious pipeline applied the Map to Reference function of Geneious Prime v.2020 (Kearse et al., 2012) to map the genome skimming data of each sample to the reference genome using Custom Sensitivity option with Allow Gaps off. Then we used the Generate Consensus Sequence function (using Trim to Reference Sequence option, and Most Common Bases for heterozygous sites) to generate sequence file of each sample, and finally saved these sequences as "fasta" files. All the alignments of this study, including the complete plastid genome sequences, SCOGs, and nuclear SNPs datasets, are deposited on Zenodo with the identifier https://doi.org/10.5281/zenodo.7215665.

Phylogenetic analysis

Plastid genome structure and gene arrangement in *Clematis* species were checked according to the method of Liu et al. (2018), and then multiple sequence alignments were done using MAFFT v.7.471 (Katoh and Standley, 2013), after removing one inverted repeat (IR) region (He et al., 2019; He et al., 2021). We used both maximum likelihood (ML) and Bayesian inference (BI) methods for phylogenetic reconstruction. ML trees were generated by RAxML v.8.2.12 (Stamatakis, 2014) under the GTR+G model with bootstrap percentages computed after 100 replicates. BI analysis was performed using MrBayes v.3.2.3 (Ronquist et al., 2012) and the best substitution model (TVM+I+G) was tested by the AIC in jModelTest v.2.1.10 (Darriba et al., 2012). Markov chain Monte Carlo (MCMC) chains run 2,000,000 generations, sampling every 100 generations. The first 25% of the trees were discarded as burn-in, and the remaining trees were used to generate the consensus tree.

For the two SCOGs datasets, we applied both concatenation- and coalescent-based methods for phylogenetic reconstruction. For the concatenation method, genes of all the datasets were concatenated. Then, we used RAxML v.8.2.12 (Stamatakis, 2014) to reconstruct phylogeny with the GTR+G model and 100 replicates of bootstrap. For the coalescent-based method, single-gene trees were reconstructed by RAxML with the parameters as above. All gene trees were then inputted in ASTRAL v.4.4.4 (Zhang et al., 2018) for species tree inference.

The nuclear SNPs matrices obtained by both GATK and Geneious pipelines had a high proportion of missing data at many loci. Therefore, we set three missing data (percentage of gaps per alignment column, Duvall et al., 2020) thresholds for each pipeline and obtained six matrices: GATK-0.4MS, GATK-0.5MS, GATK-0.6MS (40%, 50%, and 60% missing data); Geneious-0MS, Geneious-0.05MS, Geneious-0.1MS (0, 5%, and 10% missing data). We used SNP-sites v.2.5.1 (Page et al., 2016) to remove the invariant sites. Then, all matrices were analyzed using ML method implemented in RAxML v.8.2.12 with "ASC_GTRGAMMA" model (Stamatakis, 2014) and 100 bootstrap replicates.

Analysis of tree discordance

In this study, we explored the discordance among nuclear gene trees, between plastid and nuclear gene trees, and analyzed the possible biological causes. We tried to exclude factors such as sampling errors, stochastic errors, and paralogs (Zou and Ge, 2008), and tested the role of incomplete lineage sorting (ILS) and hybridization on the discordance of gene trees.

First, we examined the conflict among nuclear gene trees (the SCOG1000 dataset). In order to reduce the influence of stochastic error, gene trees with average support values more than 60 were chosen for analysis. We used Phyparts v.0.0.1 (Smith et al., 2015) to compare each nuclear gene tree with the species tree, calculated the proportion of gene trees concordant with the species tree at each node, and displayed them with pie charts. Meanwhile, to further visualize single-gene tree conflicts, we built cloud tree plots using the python package Toytree v.2.0.5 (Eaton, 2020).

The causes of nuclear gene tree conflicts were explored using a multiple species coalescent (MSC) model implemented in a simulation analysis to investigate whether ILS could be used to explain the conflict among nuclear gene trees (Yang et al., 2020; Morales-Briones et al., 2021). If the coalescent model fit the empirical gene trees well, the simulated gene trees would be consistent with the empirical gene trees, and ILS can explain the tree discordance. We used the function "sim.coaltree.sp" in the R package Phybase v.1.5 (Liu and Yu, 2010) to simulate 10,000 gene trees under the MSC model (the input coalescent species tree was constructed using the SCOG1000 dataset). Finally, we calculated the distances between each empirical gene tree and the species tree using DendroPy v.4.5.2 (Sukumaran and Holder, 2010), then showed the distance distribution between simulated gene trees and species tree using a histogram plot.

We also analyzed the causes of cyto-nuclear discordance and carried out a coalescent simulation study (Rose et al., 2021). We used the "sim.coaltree.sp" function in the R package Phybase v.1.5 (Liu and Yu, 2010) to simulate 10,000 gene trees, and then used PhyParts v.0.0.1 (Smith et al., 2015) to compare these simulated gene trees to the plastome phylogeny. If the discordant nodes are supported by a certain proportion of simulated gene trees, then it is probable that the conflict was caused by incomplete lineage sorting.

Results

Data of genome skimming, transcriptome, and draft genome

The genome skimming data size of each sample ranged from 5.07 Gbp (*C. songorica*) to 6.20 Gbp (*C. brevicaudata*), and the Q20 was 96.0%–98.9%, Q30 was 89.4%–96.9% (Supplementary Table S1). The data size of transcriptomes ranged from 5.41 Gbp (*C. viridis*) to 6.76 Gbp (*C. sibirica* Miller), and the Q20 was 97.3%–98.5%, Q30 was 93.0%–95.6%. The number of *de novo* transcripts varied from 53,429 (*C. tibetana*) to 147,758 (*C. macropetala* Ledeb.), and 37,188–114,171 transcripts were kept after removing redundancy. The N50 length of the transcripts ranged from 712 bp to 1,547 bp, and completeness of the assemblies comparing to BUSCO ranged from 54.2% (*C. reticulata* Walter) to 75.9% (*C. terniflora* DC.) (Supplementary Table S2). The size of *C. brevicaudata* genome draft was 7.81 Gbp with the contig N50 being 2,579 bp, and the number of contigs longer than 500 bp was 4,293,110 in total size of 6.19 Gbp.

Plastid genome, nuclear SCOGs, nuclear SNPs data

We acquired a total of 32 *Clematis* plastome sequences ranging from 159,284 bp (*C. reticulata*) to 159,847 bp (*C. viridis*). The number and arrangement of the plastid genes of all the *Clematis* species are identical, all contained a pair of IRs (31,023–31,082 bp.) separated by a large single copy region (79,074–79,693 bp) and a small single copy region (17,978–18,229 bp). All plastomes encoded a set of 112 genes, including 79 protein-coding genes, 29 transfer RNAs and four ribosomal RNAs (Supplementary Table S3). After removing IRa and poor alignment region, we finally obtained a matrix with aligned length of 128,149 bp for phylogenetic analysis.

For the transcriptome data, we obtained 9,900 SCOGs by homologous clusters after removing 106 organelle genes. We further discarded 3,782 genes, which were shorter than 1,000 bp, in subsequent analyses. Finally, the SCOG1000 dataset contained 6,118 genes (4,393 genes with average support value over 60), and SCOG3000 dataset contained 699 genes.

The data amount of nuclear SNPs matrix obtained by GATK and Geneious pipelines are different. The lengths of the matrices obtained by GATK pipeline are 21,767bp (GATK-0.4MS), 48,933 bp (GATK-0.5MS), and 100,223 bp (GATK-0.6MS), whereas those of the Geneious pipeline are 99,179 bp (Geneious-0MS), 375,536 bp (Geneious-0.05MS), and 2,066,289 bp (Geneious-0.1MS), respectively. The proportion of missing data in sect. *Naravelia* Prantl and sect. *Naraveliopsis* Hand.-Mazz. were significantly higher than those in other species. Because only 2.59 Gbp genome skimming data were available online for *C. fusca* Turcz. (Supplementary Table S1), high percentage of missing data was also present in its nuclear SNPs sequence.

Phylogenetic analysis

Plastid phylogeny

For the plastid genome data, except a few clades with relatively weak support values, the majority of branches received full support (Figure 2A). Sect. *Naraveliopsis* Hand.-Mazz., sect. *Atragene* (L.) DC., sect. *Naravelia* (DC.) Prantl, sect. *Cheiropsis* DC., sect. *Meclatis* (Spach) Baillon, and sect. *Fruticella* Tamura (sensu Tamura, 1995) were shown to be monophyletic. Whereas, some sections, such as sect. *Campanella* Tamura, sect. *Clematis* (sensu Tamura, 1995), and sect. *Tubulosae* Decne. were not supported in the plastid phylogeny, and species of these sections were nested together.

Phylogeny of nuclear SCOGs

Two transcriptome-based datasets (SCOG1000 and SCOG3000) yielded highly congruent phylogenies in the coalescent-based and concatenated analyses (Figure 3 and Supplementary Figure S1). All nodes in the SCOG1000 dataset using coalescent method obtained 100% support values. Whereas, in the SCOG3000 dataset, sect. *Fruticella* was not 100% supported, and the position of *C. songorica* was different in the coalescent and concatenated analyses (Supplementary Figures S1, S2). For the SCOG1000 dataset, except sect. *Campanella*, sect. *Viorna* (sensu Wang and Li, 2005) and sect. *Clematis* (sensu Wang and Li, 2005), which was shown to be polyphyletic, other sections were supported (Figure 3). Sect. *Clematis* (sensu Tamura, 1995) and sect. *Tubulosae* were both supported and tested to be sister groups

Nuclear SNPs phylogeny

The nuclear SNPs phylogenies based on the GATK pipeline were slightly different in basal branches which were insufficiently supported (Supplementary Figures S3). Among them, GATK-0.4MS dataset yielded a phylogeny which was more consistent with the trees inferred from the Geneious pipeline (Supplementary Figure 4). The resolved clades were also largely consistent with the SCOG1000 species tree (Figure 3). However, although sect. *Clematis* (sensu Tamura, 1995) and

sect. *Tubulosae* showed close relationship in the three GATK datasets, the former section was shown to be paraphyletic to the latter (Supplementary Figure S3).

The phylogenies inferred from the three datasets of Geneious pipeline were basically similar, but differed in support values (Figure 4 and Supplementary Figure S4). The Geneious-0.05MS dataset produced the most robust phylogeny, which was almost the same with the SCOG1000 species tree. Their major difference was the position of sect. *Naraveliopsis* (Figures 3, 4). Both SCOG1000 and the Geneious-0.05MS datasets had some well-supported incongruence with the plastid tree (Figure 2; Supplementary Figure S2). Because the nuclear SNPs dataset contains more samples than the SCOG1000 dataset, we discuss the phylogenetic relationships of *Clematis* mainly based on Geneious-0.05MS dataset (Figure 4). In this phylogenetic tree, ten major clades were resolved, and one section (sect. *Campanella*) in Tamura (1995) and two sections (sect. *Clematis* and sect. *Viorna*) in Wang and Li (2005) were shown to be polyphyletic. All the other sectional classification of the two systems were supported.

Gene conflict analyses

Using SCOG1000 (and average bootstrap value more than 60) dataset, high levels of gene tree discordances were detected mainly at deep nodes (Figure 5). Coalescent simulation analysis (Figure 6) showed similar pattern between empirical and simulated distance distributions, indicating that ILS alone can explain most of the gene tree conflicts. However, the contradiction between some nodes of the plastid and the nuclear species trees cannot be explained by ILS (Figure 2A). For example, species of Clade 9 (sect. *Clematis* sensu Tamura, 1995) and Clade 10 (sect. *Tubulosae*) (Figure 2B) were clustered together in the plastome phylogeny (Figure 2A), and Phyparts result showed no simulated gene trees were concordant to the empirical plastome tree. Moreover, some species of sect. *Campanella* (such as *C. rehderiana*) also showed different positions in the simulated gene trees and the plastome tree, suggesting that ILS can be excluded for explaining its cyto-nuclear discordance, and hybridization and introgression might be the main cause.

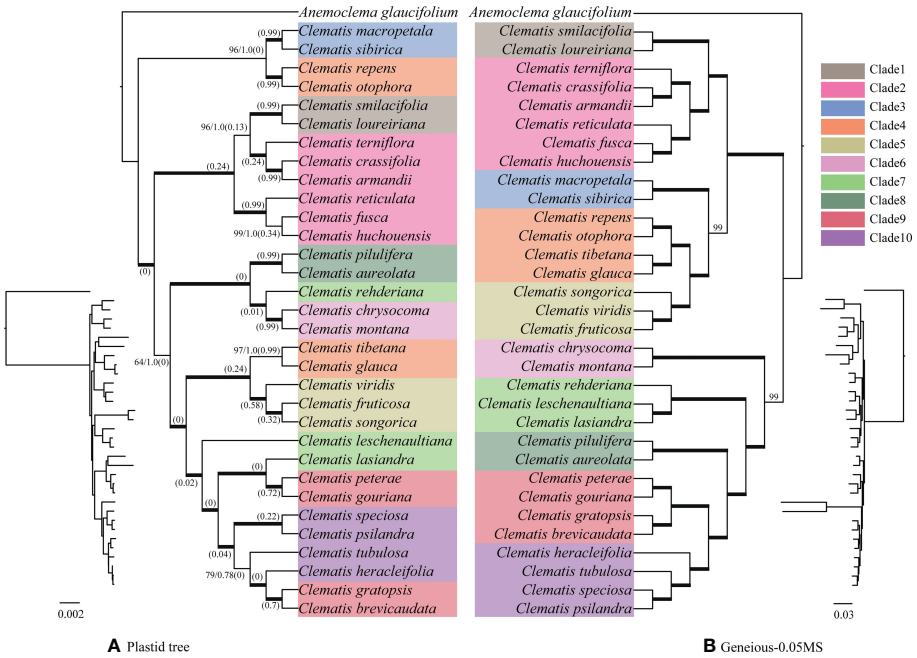


FIGURE 2
Bayesian phylogeny **(A)** of *Clematis* inferred from the plastid genome data and maximum likelihood phylogeny **(B)** inferred from nuclear SNPs of Geneious-0.05MS dataset. Cyto-nuclear conflicts are shown. In the plastid phylogeny **(A)**, bold branches show that the clades are 100% supported by both posteriori probability and ML bootstrap values. Otherwise, these two statistical values were marked on the branches. Numbers in brackets show the contribution of incomplete lineage sorting (ILS) to the conflicts between the simulated and plastid gene trees based on the multispecies coalescent model. Ten major clades were marked on the nuclear SNPs tree **(B)** with different colors. Species in plastid tree were marked with the same color with those in the nuclear SNPs tree.

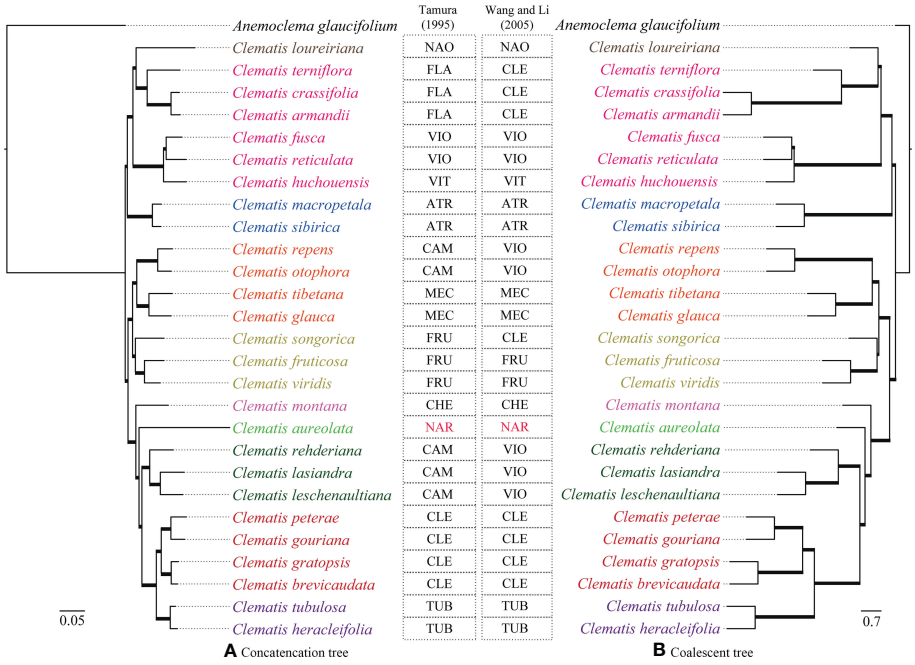


FIGURE 3
Phylogenetic trees inferred from SCOG1000 dataset by the concatenated (left) and the coalescence-based (right) methods. All clades of both trees are 100% supported and all the branches are in bold. Section abbreviations are: sect. *Atragene* (ATR), sect. *Naraveliopsis* (NAO), sect. *Clematis* (CLE), sect. *Flammula* (FLA), sect. *Viorna* (VIO), sect. *Viticella* (VIT), sect. *Campanella* (CAM), sect. *Meclatis* (MEC), sect. *Fruticella* (FRU), sect. *Cheiropsis* (CHE), sect. *Tubulosae* (TUB), *Naravelia* (NAR).

Discussion

Phylogenomic data for *Clematis*

Seed plants encompass a high level of diversity of genome size varying by more than 2,000-fold (Novák et al., 2020). Larger genomes generally contain more proportion of repeat sequences, transposable elements, and other non-transcribed low-copy sequences, while the amount of expressed genes are rather stable with about 0.03 Gbp (Kersey, 2019). For this reason, we do not need to consider the plant genome size when choosing transcriptome method for phylogenetic studies. However, when choosing genome skimming data, the size of plant genome becomes a vital issue that should be considered.

Clematis species have large genome size, which makes high-depth sequencing (10 × or more) unaffordable, and low-depth genome skimming data (less than 1 ×) of *Clematis* have been only used for assembling the plastome sequences or tandemly repeat nrDNA regions which have high copy numbers in the genome (He et al., 2021). The plastome phylogeny of *Clematis* (He et al., 2021) have better resolved the relationships within the genus than those of the Sanger sequencing data (e.g., Miikeda et al., 2006; Xie et al., 2011; Lehtonen et al., 2016). However, there were still some major clades with weak support and some clades are unexplainable taxonomically. The nrDNA sequences also failed to generate a robust tree due to insufficient phylogenetic information (He et al., 2021).

Although plastome data have been successfully used for phylogenetic reconstruction of plant taxa at almost all taxonomic levels, studies have shown that the plastome data alone may sometimes not sufficiently resolve the phylogeny of closely related species due to frequent hybridization and introgression in plants (Liu et al., 2022). Therefore, care should be taken when using plastome data alone to resolve species relationships of plant taxa. This is also the case with *Clematis*. Evidences from horticulture (Yuan et al., 2010), molecular phylogenetic studies (Lyu et al., 2021), and the present study showed that there is widespread hybridization among *Clematis* species or even between sections. In this study, transcriptome data were successfully assembled with thousands of SCOGs which robustly resolved the phylogenetic framework of *Clematis*. The SCOG1000 dataset not only fully resolved *Clematis* phylogeny but also provided a tree that corresponded well to morphological groups. The RNA-seq method is easy, fast, efficient for acquiring highly reusable nuclear genome data, independent of plant genome size (Cheon et al., 2020), and maybe the best choice for phylogenetic study of *Clematis* so far. The major problem with transcriptome method is that it cannot be successfully applied for herbarium materials. If we want to include more herbarium samples, data partitioning method should be reconsidered.

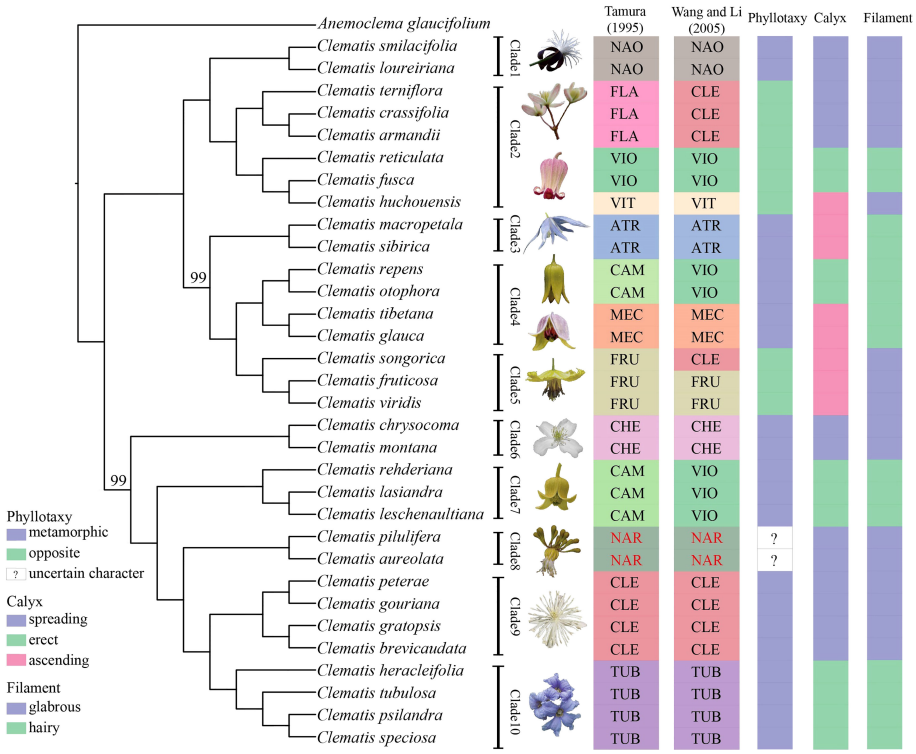


FIGURE 4
A maximum likelihood tree inferred form a nuclear SNPs dataset (Geneious-0.05MS). Two bootstrap values, which are less than 100, are marked above the branch, and all the other branches are fully supported. Section abbreviations follow Figure 3, and three important morphological characters are marked at right side of the tree.

Previous studies have used genome skimming data to obtain nuclear SNPs by mapping reads to the reference genome (Olofsson et al., 2019; Zhang et al., 2019). This study presented a further exploration of this method in *Clematis*. The phylogenies from the nuclear SNPs data by the two pipelines in this study were better resolved than the previous published nrDNA tree (He et al., 2021). Two different pipelines generated different amounts of data, and Geneious pipeline produced larger datasets than GATK pipeline. In the same way, Geneious pipeline generated more robust phylogeny which was almost the same with that reconstructed by SCOGs. Meanwhile, two herbarium samples (*C. psilandra* and *C. speciosa*) and four genome skimming data (*C. fusca*, *C. macropetala*, *C. pilulifera*, and *C. reticulata*) from other studies with lower sequencing depth clustered in the correct positions on the nuclear SNPs tree (Figure 4). Therefore, this method (especially the Geneious pipeline) is reliable and may play an important role in future phylogenetic study of *Clematis* with comprehensive sampling.

The problems of this method, however, also need to be mentioned. Because the sequencing depth is low, SNP genotyping and allele frequency estimation may be biased by those genome skimming data. So, the SNPs datasets may not be applied for population genetic analysis, such as STRUCTURE (Pritchard et al., 2000). Furthermore, this data may also not work well for analysis of reticulate evolution (such as HyDe, Blischak et al., 2018) and whole genome duplication detection (WGD, Yang et al., 2019).

Phylogenetic inferences of *Clematis*

Although previous phylogenetic studies used more samples (Xie et al., 2011; Lehtonen et al., 2016), insufficient resolution by small number of DNA regions has hindered our understanding of the evolution of *Clematis*. The plastome data took us a step forward in resolving the phylogeny of the genus (He et al., 2021). Plastome phylogeny, inferred by He et al. (2021), resolved six major clades in *Clematis*. Except a clade comprising only species of sect. *Naravelia*, all the other five clades contained three or more sections. Despite the smaller sample size of this study, all the six corresponding clades were also resolved by our plastome phylogenetic analysis. These clades (except sect. *Naravelia* clade) were difficult to be defined by morphology.

In this study, using nuclear SNPs and SCOG data, we reconstructed the first well resolved phylogenetic backbone of *Clematis*. Most of the morphologically defined sections were

10.3389/fpls.2022.1059379

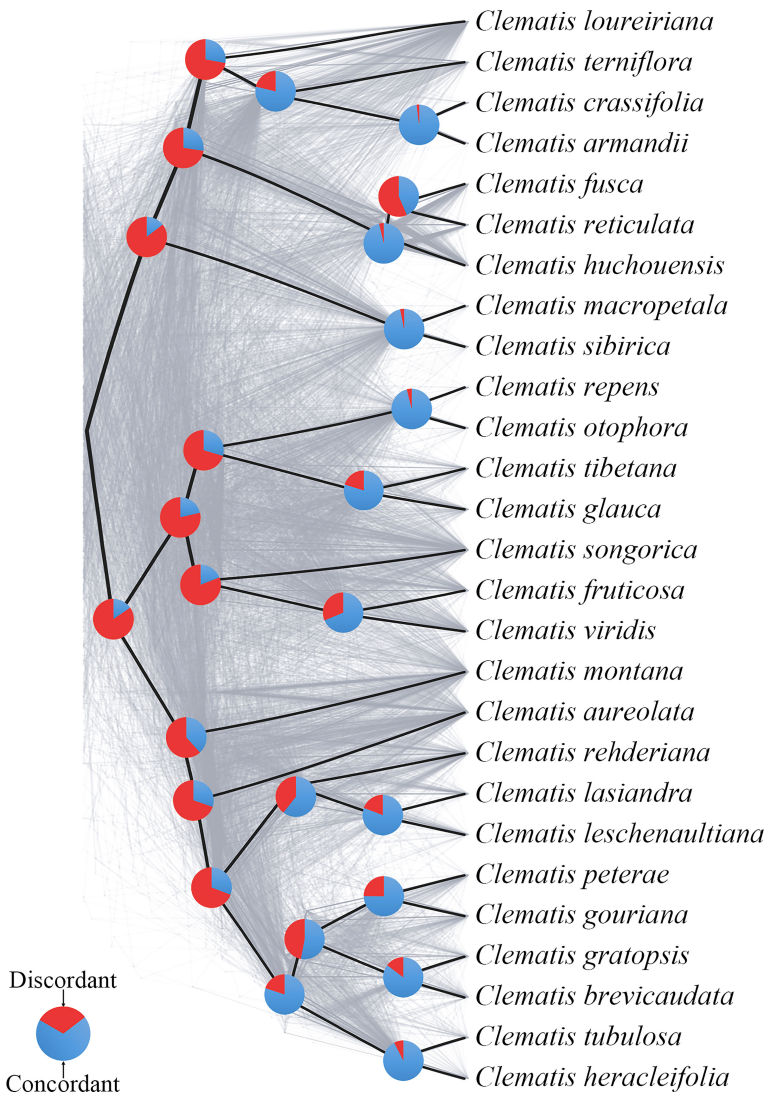


FIGURE 5
A cloud tree showing discordance among nuclear genes. The ASTRAL species tree (based on trees from SCOG1000 dataset with average bootstrap value more than 60) is in heavy black lines. All the branches are fully supported. The gray-colored trees (cloud tree) were sampled from 695 SCOGs (without missing taxa). Pie charts show the proportions of concordant and discordant topologies of gene trees comparing to the species tree.

supported. Trees inferred from the nuclear genome data (Figures 2–4) were better corresponding to morphological characters than plastome phylogeny. The Geneious-0.05MS dataset resolved ten major clades in *Clematis* (Figures 2, 4). Clade 1 represents subtropical sect. *Naraveliopsis* which has conspicuous connective projections on the anthers. Clade 2 comprises species of sect. *Flammula* DC., sect. *Viticella* DC. and sect. *Viorna* (sensu Tamura, 1995). The synapomorphy may be their type II seedlings (or opposite seedling leaves, Essig, 1991). Clade 3 represents sect. *Atragene*, which has petal-like staminodes in the flowers. Clade 4, including sect. *Meclatis* and species of sect. *Campanella* with yellow flowers and hairy filaments and anthers, is characterized by its yellow and thick sepals. Clade 5 represents sect. *Fruticella* with erect shrubby stem. Clade 6 represents sect. *Cheiropsis*, which is characterized by its flowers arising from old or hornotinous branches. Clade 7 contains some species of sect. *Campanella*. Their shared characteristics are the type I seedling (or alternate seedling leaves, Essig, 1991), erect sepals, hairy stamen filaments and glabrous anthers. Clade 8 is sect. *Naravelia* which was recognized as a distinct genus by Tamura (1995) and Wang and Li (2005). Plants of this section possess leaf tendrils and spoon-shaped petals. Clade 9 represents the narrowly defined sect. *Clematis* (sensu Tamura, 1995), which is characterized by the type I seedling, small white flowers, spreading sepals, and glabrous stamens. Clade 10 represents sect. *Tubulosae*, which is characterized by the type I seedling,

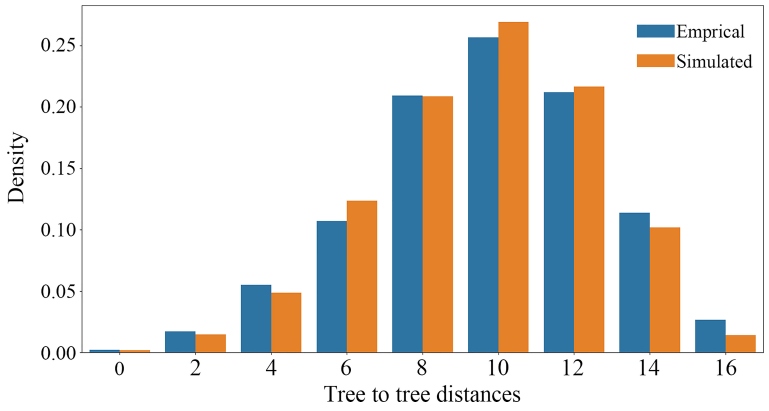


FIGURE 6
Coalescent simulations of tree-to-tree distance distributions between the ASTRAL species tree and the 4393 empirical (orange boxes) gene trees (based on trees from SCOG1000 dataset with average bootstrap value more than 60) and those from the 10,000 simulation trees (blue boxes).

ternate leaves, erect herbaceous stem, erect sepals and hairy stamens. It should be pointed out that the range of our sample was relatively narrow, and future studies with more comprehensive sampling are needed to further elucidate the phylogenetic and taxonomic problems in *Clematis*.

Using nuclear genome data, we also gained new knowledge and insights into some taxonomic issues for *Clematis* in this study. Previous studies have suggested that sect. *Campanella* may be a polyphyletic group (He et al., 2021). Our nuclear genome phylogeny confirmed that *C. repens* and *C. otophora* (in sect. *Campanella*) are more closely related to sect. *Meclatis* (clade 4, Figure 4) rather than to other sect. *Campanella* species. Both *C. repens* and *C. otophora* have yellow flowers with thick sepals which are more similar to those of the sect. *Meclatis*. Two morphologically well diverged sections, sect. *Clematis* (sensu Tamura, 1995) and sect. *Tubulosae*, have shown to be very closely related or even cannot be clearly separated by Sanger sequencing data (Xie et al., 2011; Lehtonen et al., 2016; Yan et al., 2016). They were also nested together in our plastome tree (Figure 2), but were clearly separated by our SCOG1000 data (Figure 3) and Geneious-0.05MS data (clade 9 and clade 10, Figures 2–4). The simulation results showed that this cyto-nuclear discordance may be caused by hybridization events between the two sections (Figure 2A). Hybridization events between these two morphologically diverged sections have been also confirmed by other reports, horticultural evidence, and phylogenomic analysis (Makino, 1907; Yuan et al., 2010; Lyu et al., 2021).

Similar to other studies, our results showed that all the important morphological characters emphasized by taxonomists, such as phyllotaxy, calyx, and filament hairs (Tamura, 1995), may have evolved multiple times, and it is difficult to make subgeneric classification by using a few key characters. Specifically, we emphasis that seedling morphology (phyllotaxy as in this study), highlighted by Tamura (1995), should be based on observations but not speculation. Majority number of *Clematis* species have no real observation data of seedling morphology. Seedling status of many sections (such as sect. *Naraveliopsis* and sect. *Fruticella*) proposed by Tamura (1995) are likely to be wrong (Cheng et al., 2016). Based on our observation, seedling morphology of sect. *Fruticella* (not published) should be type I (metamorphic, Essig, 1991) and similar to that of sect. *Meclatis* (rather than type II proposed by Tamura, 1995). So, before using seedling morphology for taxonomic treatment, this character needs to be studied through comprehensive observation.

Our findings also shed light on the evolutionary history of *Clematis*. Studies have shown that *Clematis* may have experienced recent species radiation during the late Neogene and the Quaternary (Xie et al., 2011; He et al., 2021). Recent species radiation may lead to severe lineage sorting when the ancestral population was large (Pamilo and Nei, 1988), and this fits well with our simulation results (Figures 5, 6). Our results demonstrated that there are extensive gene tree conflicts at early diverged nodes, which can be explained by ILS. Meanwhile, our analysis of cyto-nuclear discordance (Figure 2) suggested that there may also have been widespread interspecific hybridization events in *Clematis*, which contributed to high level of incongruence between plastid and nuclear phylogenies. From our analysis, both ILS and interspecific hybridization in *Clematis* made its classification and phylogenetic analysis very difficult, especially using small number of DNA regions or plastome data alone.

Consideration of genome partitioning selection for other plant taxa

There are several other genera in Ranunculaceae that are similar to *Clematis*, such as *Anemone* L., *Aconitum* L., and

Delphinium L. These genera have not only large genome size (https://cvalues.science.kew.org/search) but also have hundreds of wild species (Tamura, 1995). In addition, they all have no high-quality whole genome reference available and few phylogenomic studies with comprehensive sampling. Resolving the phylogenetic framework of those taxa is highly possible to encounter the same conditions with *Clematis*: ineffectiveness of Sanger sequencing data and difficulty in genome partitioning selection. Furthermore, studies have shown that the plastid genome (or regions) data alone did not work well for the phylogenetic reconstruction of those taxa (Hoot et al., 2012; Jiang et al., 2017; Hong et al., 2017; Xiang et al., 2017). Our results suggested that transcriptome method may be the first choice for solving the problem, and if the samples are not suitable for RNA extraction, Geneious pipeline presented in this study (using low-depth genome skimming data) can be tried. Although this study did not test target enrichment data, this method is also recommended if the complicated experimental procedures are acceptable to the researchers.

Genome size may be an important factor in genome partitioning selection. If the genome size of concerning taxon is small (less than 1 Gbp), genome skimming method can easily obtain high sequencing depth at an acceptable cost, and is a good choice to solve phylogenetic problems. We have tried to obtain and successfully assembled SCOGs (not published) from 6 Gbp of genome skimming data from *Epilobium* L. (Onagraceae) samples using the method of Liu et al. (2021). The genome size of *Epilobium* species is about 0.2 Gbp, and our data was up to 30 × in sequencing depth. In this case, genome skimming method is better than transcriptome and target enrichment method. Using this data, we can acquire the plastome and nuclear SCOGs data from both transcribed region and non-transcribed (intron, spacer, repetitive regions, and so on) regions, and conduct a variety of downstream analysis, such as phylogenetic reconstruction, molecular dating, hybridization analysis, and WGD detection.

Data availability statement

The data presented in the study are deposited in the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/), accession number PRJNA838588 and PRJNA776151.

Author contributions

JX, RL, and JH, analyzed the data and prepared the draft. ML, JH, JJ, and LX conducted the sample gathering. JH, JC, and LX designed the study. JX, RL, and LX wrote and revised the manuscript. All the authors contributed to the article and approved the submitted version.

Funding

This study was supported by the National Natural Science Foundation of China (grant numbers 32270223, 31670207).

Acknowledgments

We thank Ma Xin-Tang and Ban Qin, working in the Herbarium of Institute of Botany, the Chinese Academy of Sciences (PE), for kindly providing *Clematis* specimen samples. We are grateful to Dr. Xu Chao from Institute of Botany, the Chinese Academy of Sciences for extracting high quality DNAs from specimen samples for this study.

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.2022.1059379/full#supplementary-material

SUPPLEMENTARY TABLE S1
Sample information of genome skimming data in this study.

SUPPLEMENTARY TABLE S2
Sample information of transcriptome data in this study.

SUPPLEMENTARY TABLE S3
Plastid genome features of the Clematis species in this study.

SUPPLEMENTARY FIGURE 1
Phylogenetic trees constructed by concatenated and coalescence-based methods based on SCOG3000 dataset.

SUPPLEMENTARY FIGURE 2
Bayesian phylogeny **(A)** of *Clematis* inferred from the plastid genome data and maximum likelihood phylogeny **(B)** inferred from of SCOG1000 data. Cyto-nuclear discordance is shown.

SUPPLEMENTARY FIGURE 3
Maximum likelihood phylogenetic trees constructed from three nuclear SNPs data matrices obtained by GATK pipeline.

SUPPLEMENTARY FIGURE 4
Maximum likelihood phylogenetic trees constructed from two nuclear SNPs data matrices obtained by Geneious pipeline.

References

Blischak, P. D., Chifman, J., Wolfe, A. D., and Kubatko, L. S. (2018). HyDe: A Python package for genome-scale hybridization detection. *Syst. Biol.* 67, 821–829. doi: 10.1093/sysbio/syy023

Buchfink, B., Xie, C., and Huson, D. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176

Chen, N. S. (2004). Using repeat masker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinf.* 5, 4–10. doi: 10.1002/0471250953.bi0410s05

Cheng, J., Yan, S. X., Liu, H. J., Lin, L. L., Li, J. Y., Liao, S., et al. (2016). Reconsidering the phyllotaxy significance of seedlings in *Clematis*. *Phytotaxa* 265, 131–138. doi: 10.11646/phytotaxa.265.2.4

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

Cheon, S., Zhang, J., Park, C., and Teeling, E. (2020). Is phylotranscriptomics as reliable as phylogenomics? *Mol. Biol. Evol.* 37, 3672–3683. doi: 10.1093/molbev/msaa181

Darriba, D., Taboada, G., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* 9, 772. doi: 10.1038/nmeth.2109

Dodsworth, S. (2015). Genome skimming for next-generation biodiversity analysis. *Trends Plant Sci.* 20, 525–527. doi: 10.1016/j.tplants.2015.06.012

Drezen, E., Rizk, G., Chikhi, R., Deltel, C., Lemaitre, C., Peterlongo, P., et al. (2014). GATB: genome assembly & analysis tool box. *Bioinformatics* 30, 2959–2961. doi: 10.1093/bioinformatics/btu406

Duvall, M. R., Burke, S. V., and Clark, D. C. (2020). Plastome phylogenomics of poaceae: Alternate topologies depend on alignment gaps. *Bot. J. Linn. Soc* 192, 9–20. doi: 10.1093/botlinnean/boz060

Eaton, D. A. R. (2020). Toytree: A minimalist tree visualization and manipulation library for Python. *Methods Ecol. Evol.* 11, 187–191. doi: 10.1111/2041-210X.13313

Essig, F. B. (1991). Seedling morphology in *Clematis* (Ranunculaceae) and its taxonomic implications. *SIDA* 1991, 377–390.

Fonseca, L. H. M., and Lohmann, L. G. (2020). Exploring the potential of nuclear and mitochondrial sequencing data generated through genome-skimming for plant phylogenetics: A case study from a clade of neotropical lianas. *J. Syst. Evol.* 58, 18–32. doi: 10.1111/jse.12533

Fu, L. M., Niu, B. F., Zhu, Z. W., Wu, S. T., and Li, W. Z. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Grey-Wilson, C. (2000). *Clematis, the genus* (Portland, OR: Timber Press).

He, J., Lyu, R. D., Luo, Y. K., Lin, L. L., Yao, M., Xiao, J. M., et al. (2021). An updated phylogenetic and biogeographic analysis based on genome skimming data reveals convergent evolution of shrubby habit in *Clematis* in the pliocene and pleistocene. *Mol. Phylogenet. Evol.* 164, 107259. doi: 10.1016/j.ympev.2021.107259

He, J., Lyu, R. D., Luo, Y. K., Xiao, J. M., Xie, L., Wen, J., et al. (2022). A phylotranscriptome study using silica gel-dried leaf tissues produces an undated robust phylogeny of ranunculaceae. *Mol. Phylogenet. Evol.* 174, 107545. doi: 10.1016/j.ympev.2022.107545

He, J., Yao, M., Lyu, R. D., Lin, L. L., Liu, H. J., Pei, L. Y., et al. (2019). Structural variation of the complete chloroplast genome and plastid phylogenomics of the genus *Asteropyrum* (Ranunculaceae). *Sci. Rep.* 9, 1–13. doi: 10.1038/s41598-019-51601-2

Hong, Y., Luo, Y., Gao, Q., Ren, C., Yuan, Q., and Yang, Q. E. (2017). Phylogeny and reclassification of *Aconitum* subgenus *Lycoctonum* (Ranunculaceae). *PLoS One* 12, e0171038. doi: 10.1371/journal.pone.0171038

Hoot, S. B., Meyer, K. M., and Manning, J. C. (2012). Phylogeny and reclassification of *Anemone* (Ranunculaceae), with an emphasis on austral species. *Syst. Bot.* 37, 139–152. doi: 10.1600/036364412X616729

Huang, D. I., and Cronk, Q. C. B. (2015). Plann: A command–line application for annotating plastome sequences. *Appl. Plant Sci.* 3, 1500026. doi: 10.3732/apps.1500026

Jiang, N., Zhou, Z., Yang, J. B., Zhang, S. D., Guan, K. Y., Tan, Y. H., et al. (2017). Phylogenetic reassessment of tribe anemoneae (Ranunculaceae): Non-monophyly of *Anemone s. l.* revealed by plastid datasets. *PLoS One* 12, e0174792. doi: 10.1371/journal.pone.0174792

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

Johnson, M. (1997). *Släktet klematis* (Södertälje: Magnus Johnson Plantskola).

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

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

Kersey, P. J. (2019). Plant genome sequences: Past, present, future. *Curr. Opin. Plant Biol.* 48, 1–8. doi: 10.1016/j.pbi.2018.11.001

Khan, G., Nolzen, J., Schepker, H., and Albach, D. C. (2021). Incongruent phylogenies and their implications for the study of diversification, taxonomy, and genome size evolution of *Rhododendron*. *Am. J. Bot.* 108, 1957–1981. doi: 10.1002/ajb2.1747

Kress, W. J., Soltis, D. E., Kersey, P. J., and Soltis, P. S. (2022). Green plant genomes: What we know in an era of rapidly expanding opportunities. *Proc. Natl. Acad. Sci. U.S.A.* 119, e2115640118. doi: 10.1073/pnas.2115640118

Lechner, M., Findeiß, S., Steiner, L., Marz, M., Stadler, P. F., and Prohaska, S. J. (2011). Proteinortho: detection of (co-) orthologs in large-scale analysis. *BMC Bioinf.* 12, 1–9. doi: 10.1186/1471-2105-12-124

Lee-Yaw, J. A., Grassa, C. J., Joly, S., Andrew, R. L., and Rieseberg, L. H. (2019). An evaluation of alternative explanations for widespread cytonuclear discordance in annual sunflowers (*Helianthus*). *New Phytol.* 221, 515–526. doi: 10.1111/nph.15386

Lehtonen, S., Christenhusz, M. J. M., and Falck, D. (2016). Sensitive phylogenetics of *Clematis* and its position in ranunculaceae. *Bot. J. Linn. Soc* 182, 825–867. doi: 10.1111/boj.12477

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* 1303, 3997. doi: 10.48550/arXiv.1303.3997

Li, H. T., Luo, Y., Gan, L., Ma, P. F., Gao, L. M., Yang, J. B., et al. (2021). Plastid phylogenomic insights into relationships of all flowering plant families. *BMC Biol.* 19, 232. doi: 10.1186/s12915-021-01166-2

Liu, H. J., He, J., Ding, C. H., Lyu, R. D., Pei, L. Y., Cheng, J., et al. (2018). Comparative analysis of complete chloroplast genomes of *Anemoclema*, *Anemone*, *Pulsatilla*, and *Hepatica* revealing structural variations among genera in tribe anemoneae (Ranunculaceae). *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01097

Liu, B. B., Hong, D. Y., Zhou, S. L., Xu, C., Dong, W. P., Johnson, G., et al. (2019). Phylogenomic analyses of the *Photinia* complex support the recognition of a new genus *Phippsiomeles* and the resurrection of a redefined *Stranvaesia* in maleae (Rosaceae). *J. Syst. Evol.* 57, 678–694. doi: 10.1111/jse.12542

Liu, B. B., Ma, Z. Y., Ren, C., Hodel, R. G. J., Sun, M., Liu, X. Q., et al. (2021). Capturing single-copy nuclear genes, organellar genomes, and nuclear ribosomal DNA from deep genome skimming data for plant phylogenetics: A case study in vitaceae. *J. Syst. Evol.* 59, 1124–1138. doi: 10.1111/jse.12806

Liu, B. B., Ren, C., Kwak, M., Hodel, R. G. J., Xu, C., He, J., et al. (2022). Phylogenomic conflict analyses in the apple genus *Malus s. l.* reveal widespread hybridization and allopolyploidy driving diversification, with insights into the complex biogeographic history in the northern hemisphere. *J. Integr. Plant Biol.* 64, 1020–1043. doi: 10.1111/jipb.13246

Liu, L., and Yu, L. (2010). Phybase: An r package for species tree analysis. *Bioinformatics* 26, 962–963. doi: 10.1093/bioinformatics/btq062

Li, J. L., Wang, S., Yu, J., Wang, L., and Zhou, S. L. (2013). A modified CTAB protocol for plant DNA extraction. *Chin. Bull. Bot.* 48, 72–78. doi: 10.3724/SP.J.1259.2013.00072

Li, H. T., Yi, T. S., Gao, L. M., Ma, P. F., Zhang, T., Yang, J. B., et al. (2019). Origin of angiosperms and the puzzle of the Jurassic gap. *Nat. Plants* 5, 461–470. doi: 10.1038/s41477-019-0421-0

Lyu, R. D., He, J., Luo, Y. K., Lin, L. L., Yao, M., Cheng, J., et al. (2021). Natural hybrid origin of the controversial "species" *Clematis* × *pinnata* (Ranunculaceae) based on multidisciplinary evidence. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.745988

Mai, U., and Mirarab, S. (2018). TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics* 19, 272. doi: 10.1186/s12864-018-4620-2

Makino, T. (1907). Observations on the flora of Japan. *Bot. Magaz. (Tokyo)* 21, 86–88.

Marks, R. A., Hotaling, S., Frandsen, P. B., and VanBuren, R. (2021). Representation and participation across 20 years of plant genome sequencing. *Nat. Plants* 7, 1571–1578. doi: 10.1038/s41477-021-01031-8

McKain, M. R., Johnson, M. G., Uribe-Convers, S., Eaton, D., and Yang, Y. (2018). Practical considerations for plant phylogenomics. *Appl. Plant Sci.* 6, e1038. doi: 10.1002/aps3.1038

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

Miikeda, O., Kita, K., Handa, T., and Yukawa, T. (2006). Phylogenetic relationships of *Clematis* (Ranunculaceae) based on chloroplast and nuclear DNA sequences. *Bot. J. Linn. Soc* 152, 153–168. doi: 10.1111/j.1095-8339.2006.00551.x

Morales-Briones, D. F., Kadereit, G., Tefarikis, D. T., Moore, M. J., Smith, S. A., Brockington, S. F., et al. (2021). Disentangling sources of gene tree discordance in phylogenomic data sets: Testing ancient hybridizations in amaranthaceae s. l. *Syst. Biol.* 70, 219–235. doi: 10.1093/sysbio/syaa066

Novák, P., Guignard, M. S., Neumann, P., Kelly, L. J., Mlinarec, J., Koblížková, A., et al. (2020). Repeat-sequence turnover shifts fundamentally in species with large genomes. *Nat. Plants* 6, 1325–1329. doi: 10.1038/s41477-020-00785-x

Olofsson, J. K., Cantera, I., Van de Paer, C., Hong-Wa, C., Zedane, L., Dunning, L. T., et al. (2019). Phylogenomics using low-depth whole genome sequencing: A case study with the olive tribe. *Mol. Ecol. Resour.* 19, 877–892. doi: 10.1111/1755-0998.13016

One Thousand Plant Transcriptomes Initiative (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574, 679–685. doi: 10.1038/s41586-019-1693-2

Page, A. J., Taylor, B., Delaney, A. J., Soares, J., Seemann, T., Keane, J. A., et al. (2016). SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb. Genom.* 2, e000056. doi: 10.1099/mgen.0.000056

Pamilo, P., and Nei, M. (1988). Relationships between gene trees and species trees. *Mol. Biol. Evol.* 5, 568–583. doi: 10.1093/oxfordjournals.molbev.a040517

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959. doi: 10.1093/genetics/155.2.945

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., et al. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. doi: 10.1093/sysbio/sys029

Rose, J. P., Toledo, C. A., Lemmon, E. M., Lemmon, A. R., and Sytsma, K. J. (2021). Out of sight, out of mind: widespread nuclear and plastid-nuclear discordance in the flowering plant genus *Polemonium* (Polemoniaceae) suggests widespread historical gene flow despite limited nuclear signal. *Syst. Biol.* 70, 162–180. doi: 10.1093/sysbio/syaa049

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Smith, S. A., Moore, M. J., Brown, J. W., and Yang, Y. (2015). Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evol. Biol.* 15, 150. doi: 10.1186/s12862-015-0423-0

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

Stull, G. W., Soltis, P. S., Soltis, D. E., Gitzendanner, M. A., and Smith, S. A. (2020). Nuclear phylogenomic analyses of asterids conflict with plastome trees and support novel relationships among major lineages. *Am. J. Bot.* 107, 790–805. doi: 10.1002/ajb2.1468

Su, C., Duan, L., Liu, P., Liu, J., Chang, Z., and Wen, J. (2021). Chloroplast phylogenomics and character evolution of eastern Asian *Astragalus* (Leguminosae): Tackling the phylogenetic structure of the largest genus of flowering plants in Asia. *Mol. Phylogenet. Evol.* 156, 107025. doi: 10.1016/j.ympev.2020.107025

Sukumaran, J., and Holder, M. T. (2010). DendroPy: A Python library for phylogenetic computing. *Bioinformatics* 26, 1569–1571. doi: 10.1093/bioinformatics/btq228

Tamura, M. (1995). "Clematis l," in *Die natürlichen pflanzenfamilien*, vol. 17a . Ed. P. Hiepko (Berlin: Duncker und Humbolt), 368–387.

Thode, V. A., Lohmann, L. G., and Sanmartín, I. (2020). Evaluating character partitioning and molecular models in plastid phylogenomics at low taxonomic levels: A case study using *Amphilophium* (Bignonieae, bignoniaceae). *J. Syst. Evol.* 58, 1071–1089. doi: 10.1111/jse.12579

Valcárcel, V., and Wen, J. (2019). Chloroplast phylogenomic data support Eocene amphi-pacific early radiation for the Asian palmate core araliaceae. *J. Syst. Evol.* 57, 547–560. doi: 10.1111/jse.12522

Vargas, O. M., Heuertz, M., Smith, S. A., and Dick, C. W. (2019). Target sequence capture in the Brazil nut family (Lecythidaceae): Marker selection and *in silico* capture from genome skimming data. *Mol. Phylogenet. Evol.* 135, 98–104. doi: 10.1016/j.ympev.2019.02.020

Wang, W. T., and Bartholomew, B. (2001). "Clematis l," in *Flora of China*, vol. 6 . Eds. C. Y. Wu and P. Raven (Beijing: Science Press, St. Louis, Missouri Botanical Garden Press), 333–386.

Wang, W. T., and Li, L. Q. (2005). A new system of classification of the genus *Clematis* (Ranunculaceae). *Acta Phytotax. Sin.* 43, 431–488. doi: 10.1360/aps040130

Wang, Y. B., Liu, B. B., Nie, Z. L., Chen, H. F., Chen, F. J., Figlar, R. B., et al. (2020). Major clades and a revised classification of *Magnolia* and magnoliaceae based on whole plastid genome sequences *via* genome skimming. *J. Syst. Evol.* 58, 673–695. doi: 10.1111/jse.12588

Watson, L. E., Siniscalchi, C. M., and Mandel, J. (2020). Phylogenomics of the hyperdiverse daisy tribes: Anthemideae, astereae, calenduleae, gnaphalieae, and senecioneae. *J. Syst. Evol.* 58, 841–852. doi: 10.1111/jse.12698

Wen, J., Harris, A. J., Ickert-Bond, S. M., Dikow, R., Wurdack, K., and Zimmer, E. A. (2017). Developing integrative systematics in the informatics and genomic era, and calling for a global biodiversity cyberbank. *J. Syst. Evol.* 55, 308–321. doi: 10.1111/jse.12270

Wen, J., Harris, A. J., Kalburgi, Y., Zhang, N., Xu, Y., Zheng, W., et al. (2018). Chloroplast phylogenomics of the new world grape species (*Vitis*, vitaceae). *J. Syst. Evol.* 56, 297–308. doi: 10.1111/jse.12447

Wikström, N., Bremer, B., and Rydin, C. (2020). Conflicting phylogenetic signals in genomic data of the coffee family (Rubiaceae). *J. Syst. Evol.* 58, 440–460. doi: 10.1111/jse.12566

Xiang, K. L., Aytaç, Z., Liu, Y., Espinosa, F., Jabbour, F., Byng, J. W., et al. (2017). Recircumscription of *Delphinium* subg. *Delphinium* (Ranunculaceae) and implications for its biogeography. *Taxon* 66, 554–566. doi: 10.12705/663.3

Xie, L., Wen, J., and Li, L. Q. (2011). Phylogenetic analyses of *Clematis* (Ranunculaceae) based on sequences of nuclear ribosomal ITS and three plastid regions. *Syst. Bot.* 36, 907–921. doi: 10.1600/036364411X604921

Yang, Y., Li, Y., Chen, Q., Sun, Y., and Lu, Z. (2019). WGDdetector: A pipeline for detecting whole genome duplication events using the genome or transcriptome annotations. *BMC Bioinf.* 20, 1–6. doi: 10.1186/s12859-019-2670-3

Yang, Y. Z., Sun, P. C., Lv, L. K., Wang, D. L., Ru, D. F., Li, Y., et al. (2020). Prickly waterlily and rigid hornwort genomes shed light on early angiosperm evolution. *Nat. Plants* 6, 215–222. doi: 10.1038/s41477-020-0594-6

Yan, S. X., Liu, H. J., Lin, L. L., Liao, S., Li, J. Y., Pei, L. Y., et al. (2016). Taxonomic status of *Clematis acerifolia* var. *elobata*, based on molecular evidence. *Phytotaxa* 268, 209–219. doi: 10.11646/phytotaxa.268.3.5

Yuan, T., Wang, L. Y., and Roh, M. S. (2010). Confirmation of *Clematis* hybrids using molecular markers. *Sci. Hortic.* 125, 136–145. doi: 10.1016/j.scienta.2010.03.005

Yu, X. Q., Yang, D., Guo, C., and Gao, L. M. (2018). Plant phylogenomics based on genome-partitioning strategies: Progress and prospects. *Plant Divers.* 40, 158–164. doi: 10.1016/j.pld.2018.06.005

Zhai, W., Duan, X., Zhang, R., Guo, C., Li, L., Xu, G., et al. (2019). Chloroplast genomic data provide new and robust insights into the phylogeny and evolution of the ranunculaceae. *Mol. Phylogenet. Evol.* 135, 12–21. doi: 10.1016/j.ympev.2019.02.024

Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinf.* 19, 15–30. doi: 10.1186/s12859-018-2129-y

Zhang, R., Wang, Y. H., Jin, J. J., Stull, G. W., Bruneau, A., Cardoso, D., et al. (2020). Exploration of plastid phylogenomic conflict yields new insights into the deep relationships of leguminosae. *Syst. Biol.* 69, 613–622. doi: 10.1093/sysbio/syaa013

Zhang, B. W., Xu, L. L., Li, N., Yan, P. C., Jiang, X. H., Woeste, K. E., et al. (2019). Phylogenomics reveals an ancient hybrid origin of the Persian walnut. *Mol. Biol. Evol.* 36, 2451–2461. doi: 10.1093/molbev/msz112

Zhao, D. N., Ren, C. Q., and Zhang, J. Q. (2021). Can plastome data resolve recent radiations? *Rhodiola* (Crassulaceae) as a case study. *Bot. J. Linn. Soc* 197, 513–526. doi: 10.1093/botlinnean/boab035

Zimmer, E. A., and Wen, J. (2015). Using nuclear gene data for plant phylogenetics: Progress and prospects II. next-gen approaches. *J. Syst. Evol.* 53, 371–379. doi: 10.1111/jse.12174

Zou, X. H., and Ge, S. (2008). Conflicting gene trees and phylogenomics. *J. Syst. Evol.* 46, 795–807. doi: 10.3724/SP.J.1002.2008.08081

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 23 November 2022
DOI 10.3389/fpls.2022.1003368

OPEN ACCESS

EDITED BY
JianJun Jin,
Columbia University, United States

REVIEWED BY
Kunli Xiang,
Institute of Botany (CAS), China
Isabel Marques,
University of Lisbon,
Portugal
Haidong Yan,
University of Georgia, United States

*CORRESPONDENCE
Wenpan Dong
wpdong@bjfu.edu.cn
Luqi Huang
huangluqi01@126.com
Lanping Guo
glp01@126.com

†These authors have contributed equally to this work

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

RECEIVED 22 August 2022
ACCEPTED 07 November 2022
PUBLISHED 23 November 2022

CITATION
Wang Y, Sun J, Qiao P, Wang J, Wang M, Du Y, Xiong F, Luo J, Yuan Q, Dong W, Huang L and Guo L (2022) Evolutionary history of genus *Coptis* and its dynamic changes in the potential suitable distribution area. *Front. Plant Sci.* 13:1003368. doi: 10.3389/fpls.2022.1003368

Evolutionary history of genus *Coptis* and its dynamic changes in the potential suitable distribution area

Yiheng Wang[1,2†], Jiahui Sun[1†], Ping Qiao[1†], Jingyi Wang[1], Mengli Wang[1], Yongxi Du[1,2], Feng Xiong[1,2], Jun Luo[3], Qingjun Yuan[1], Wenpan Dong[4*], Luqi Huang[1*] and Lanping Guo[1,2*]

[1]State Key Laboratory Breeding Base of Dao-di Herbs, National Resource Center for Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing, China, [2]Key Laboratory of Biology and Cultivation of Herb Medicine, Ministry of Agriculture and Rural Affairs, Beijing, China, [3]Kunming Xishan Forestry and Grassland Comprehensive Service Center, Kunming, China, [4]Laboratory of Systematic Evolution and Biogeography of Woody Plants, School of Ecology and Nature Conservation, Beijing Forestry University, Beijing, China

The genus *Coptis* belongs to the Ranunculaceae family, containing 15 recognized species highly diverse in morphology. It is a conspicuous taxon with special evolutionary position, distribution pattern and medicinal value, which makes it to be of great research and conservation significance. In order to better understand the evolutionary dynamics of *Coptis* and promote more practical conservation measures, we performed plastome sequencing and used the sequencing data in combination with worldwide occurrence data of *Coptis* to estimate genetic diversity and divergence times, rebuild biogeographic history and predict its potential suitable distribution area. The average nucleotide diversity of *Coptis* was 0.0067 and the hotspot regions with the highest hypermutation levels were located in the *ycf1* gene. *Coptis* is most likely to have originated in North America and Japanese archipelago and has a typical Eastern Asian and North American disjunct distribution pattern, while the species diversity center is located in Mid-West China and Japan. The crown age of the genus is estimated at around 8.49 Mya. The most suitable climatic conditions for *Coptis* were as follows: precipitation of driest quarter > 25.5 mm, annual precipitation > 844.9 mm and annual mean temperature -3.1 to 19 °C. The global and China suitable area shows an upward trend in the future when emission of greenhouse gases is well controlled, but the area, especially in China, decreases significantly without greenhouse gas policy interventions. The results of this study provide a comprehensive insight into the *Coptis* evolutionary dynamics and will facilitate future conservation efforts.

KEYWORDS

***Coptis*, plastid genetic diversity, divergence time, biogeographic history, potential suitable distribution, phylogenetic relationship**

Introduction

Understanding the evolutionary dynamics of living organisms is central to characterize biodiversity on Earth (Dong et al., 2022). The study of taxa with a significant phylogenetic position is a research hotspot in evolutionary biology. Ranunculaceae, located in the early-diverged of eudicot, has fascinated botanists for decades due to its fantastic features in diversification, evolution, and phylogeny (Zeng et al., 2014; Wei et al., 2019). In addition, intercontinental disjunction distribution is another research hotspot that has attracted considerable attention from botanists and biogeographers (Duan et al., 2020; Song et al., 2020; Nge et al., 2022). Understanding the past disjunct distribution pattern, timing, and the associated drivers is also a critical step in elucidating a clear evolutionary story. The Eastern Asian and North American (EA-NA) distribution is one of the most well-known disjunct distribution patterns, and the Bering land bridge probably provided opportunities for floristic exchange (Feng et al., 2021; Ye et al., 2022). Distribution of species was significantly shaped by environmental conditions, including temperature and precipitation. As a result, modeling species distribution has become popular in conservation, ecology, biogeography and evolution studies (Li et al., 2020; Huang et al., 2020).

Coptis, belongs to the Ranunculaceae family, which contains 15 recognized species that are highly diverse in floral morphology, but most species are extremely endangered, and disjunctly distributed from the warm-temperate to the boreal zone of East Asia and North America (Tamura, 1995). The species of this genus are also important medicinal plants worldwide. 'Huanglian', the rhizome of *Coptis* plants, has been clinically used as an antiviral, antimicrobial and anti-inflammatory agent for thousands of years (He et al., 2014). Currently, as the COVID-19 pandemic sweeps the world, 'Huanglian' shows the potential for fighting against the epidemic (Ma et al., 2020; Xu et al., 2022). In general, *Coptis* has a special evolutionary position, typical disjunct pattern and important medicinal value, which makes it to be of great research and conservation significance. To better characterize this special genus, an updated understanding of its evolutionary history, through a robust study on genetic variation, phylogeny, divergence time, and past disjunct distribution pattern is required. In addition, it is also necessary to develop more practical conservation measures, and predict its potential suitable distribution area from present to future.

As a well-known Chinese herbal medicine, the synthetic pathways of alkaloids derived from *Coptis* species, and their pharmacological and pharmacodynamic characteristics have been extensively studied (Minami et al., 2007; Yan et al., 2008; Chen et al., 2021; Liu et al., 2021; Yang et al., 2021). A few studies focused on phylogenetic reconstruction, distribution patterns assessment, divergence time estimation, and prediction of suitable distribution area (He et al., 2014; Xiang et al., 2016; Xiang et al., 2018; Li et al., 2020; Wang et al., 2020). Adequate sampling and sufficient variable molecular information are necessary for a reliable evolutionary process and prediction of the distribution. However, these studies only use a few DNA markers (*rbcL*, *matK*, *trnL-F*, *trnH-psbA*, ITS, etc.) which have limited variable information or the sample size of the studies was small that not all *Coptis* species from multiple distribution areas were included. Therefore, we expanded sampling to cover almost all the species of *Coptis* and used more informative markers to perform evolutionary research. In the meanwhile, we added more occurrence records to cover the worldwide distribution of *Coptis* for precise prediction of its suitable areas.

With the development of next-generation sequencing technologies, plastome genome sequences can be obtained more efficiently by directly sequencing total genomic DNA and *de novo* assembling whole plastid genomes (Sun et al., 2021). Due to its stable structure, rare recombination, moderate evolution rate, and largely uniparental inheritance, plastomes have been extensively used to reveal the evolutionary dynamics, such as phylogeny, phylogeography, demographic history and species diversity estimation (Aguirre Planter et al., 2020; Dong et al., 2021a; Dong et al., 2021b; Wang Y. et al., 2021; Wang Z. et al., 2021).

In this study, we newly sequenced, assembled eleven plastomes and analyzed twelve plastomes of *Coptis* (one of which was downloaded from GenBank). Combined with the geographical and climatic data of *Coptis* global distribution area, we aimed to (1) evaluate the plastome variation and identify the most variable regions, (2) reconstruct the phylogenetic relationship and estimate divergence times, (3) rebuild the biogeographic history and infer the formation of their disjunct pattern, and (4) predict its potential suitable distribution area. By combining these aspects, we aim to elucidate the evolutionary dynamics of this important genus and propose more reasonable protection measures.

Material and methods

Plant material and DNA extraction

Eleven specimens of genus *Coptis* were obtained from the herbarium of PE (Herbarium, Institute of Botany, CAS, Beijing, China) and CMMI (Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing, China) (Table S1). The sequence of *C. japonica* was downloaded from GenBank (NC054329). Total genomic DNA was extracted from specimens using a modified cetyl trimethyl ammonium bromide (CTAB) method and purified with the Wizard DNA clean-up kit (Promega Corporation, Madison, WI, USA) (Li et al., 2013). DNA quality was assessed by spectrophotometry and stored at the -20°C.

Plastome sequencing, assembly, and annotation

After fragmenting into 300-350 bp fragments by sonication, a pair-end library was constructed using the NEBNext Ultra™ DNA library prep kit (New England Biolabs, Ipswich, MA, USA). Pair-end (PE150) sequencing of 11 accessions was performed on the Illumina HiSeq XTen platform at Novogene Co., Ltd (Beijing, China). The raw data of the PE150 sequencing were filtered using the Trimmomatic 0.39 software (with settings: ILLUMINACLIP : TruSeq3-PE. fa:2:30:10:1:true LEADING:20 TRAILING:20 SLIDINGWINDOW:4:15) (Bolger et al., 2014). *De novo* assembly of the high-quality reads was performed using the Getorganelle v1.7.5 software with the suggested settings: -R 15 and -k 85,105 (Jin et al., 2020). Ultimately, all reads were mapped to the assembled plastome sequence to verify the assembly accuracy in Geneious 8.1 software (Biomatters Ltd., Auckland, New Zealand) as a double-check process. Gene annotation was performed using the online platform CPGAVAS2 with default settings and manual checking in Sequin to avoid missing or incorrectly annotated genes (Shi et al., 2019). The circle maps of the plastomes were plotted using the online program Chloroplot (https://irscope.shinyapps.io/Chloroplot/).

Comparison of plastomes

Twelve plastomes were aligned using the multiple alignment software MAFFT online (https://mafft.cbrc.jp/alignment/server/) and manually adjusted using Se-al 2.0. Comparison of the whole plastomes of *Coptis* species was performed using the mVISTA program (http://genome.lbl.gov/vista/mvista/submit.shtml) in the Shuffle-LAGAN mode with *C. aspleniifolia* as a reference. Additionally, the nucleotide diversity (Pi, π) and Indels were calculated based on a 500-bp sliding window using the DnaSP v5.10 software (Librado and Rozas, 2009). Circos analysis was performed on the indel and nucleotide diversity data using the OmicStudio tools (https://www.omicstudio.cn/tool/) to visualize the hotspot region. A line graph was plotted to show the hyper-mutation region in detail (Gu et al., 2014).

Phylogeny, biogeography and divergence time estimation

A total of 19 plastome sequences were used to reconstruct the phylogeny, including the *Coptis* species and seven outgroup plastomes from GenBank (Table S1). All genome sequences were aligned using the MAFFT software and ambiguous regions were trimmed by the Gblocks 0.91b program (Castresana, 2000). The program ModelFinder was used to select the best-fit model according to the Bayesian information criterion (Kalyaanamoorthy et al., 2017). The maximum likelihood tree was inferred using IQ-TREE with the TVM+F+R3 model and 5,000 ultrafast bootstraps in PhyloSuite (Nguyen et al., 2015; Zhang et al., 2020). Bayesian Inference phylogenies were inferred using MrBayes 3.2.6 (Ronquist et al., 2012) under GTR+I+G+F model (12 parallel runs, 500,000 generations), in which the initial 25% of sampled data were discarded as burn-in. Trees were visualized in FigTree v1.3.1.

Divergence time estimation was performed using a relaxed log normal clock model in the BEAST v2.6.6 platform, using the GTR substitution model and a speciation Yule Process tree prior (Bouckaert et al., 2019). The Markov chain Monte Carlo chains (MCMC) were run for 900 million generations and sampled every 1,000 generations with a sampling frequency of 1,000 generations. Secondary calibration points for dating are listed in Table S2 according to previous studies (Xiang et al., 2018; Wei et al., 2019). The adequate effective sample size values (ESS > 200) were checked in Tracer 1.6. After a burn-in of 25%, a maximum clade credibility (MCC) tree with 95% highest posterior density intervals on each node was calculated using TreeAnnotator 2.1.3 and displayed in FigTree v1.3.1.

Based on the present distribution, we delimited four biogeographical areas: A, China mainland; B, Taiwan island; C, Japanese archipelago and a part of the Russian Far East; D, North America. We estimated ancestral distributions using the R package BioGeoBEARS implemented in Reconstruct Ancestral State in Phylogenies (RASP 4.0) (Yu Y. et al., 2020). The Dispersal Extinction Cladogenesis model with the jump dispersal parameter (DEC+J) was taken as the best model according to the model test in RASP (Table S3). In addition, we showed only the most likely status (MLS) for nodes where dispersal or vicariance had occurred.

Coptis occurrences and species diversity

The worldwide occurrence records of *Coptis* species were collected from the Global Biodiversity Information Facility (GBIF; occurrence download https://doi.org/10.15468/dl.zqzb7b) and the National Plant Specimen Resource Center (NSII; http://www.nsii.org.cn/). Two-step approach for distribution records quality control was carried out. First, checking the species name in The Plant List (http://www.theplantlist.org/), removed unresolved or incorrect species records and corrected synonyms species. Second, generating the respective distribution maps for each species in ArcGIS 10.8, manual check the points inconsistent with the flora description and verified the species credibility. After removing or revising distribution points, 9,182 distribution records of *Coptis* species were obtained (Table S4). The global land area was divided into 2° grid cell by ArcGIS and the number of species per grid were counted to determine the species diversity of *Coptis*. A density map was plotted to visualize the distribution pattern of *Coptis* species diversity.

Climatic variables and distribution modeling

In order to reduce sampling deviation in suitable habitats predicted by the MaxEnt model, we performed spatial rarefying of the data obtained in the previous step on a 10 km resolution using the SDM toolbox v2.5 and 2,898 records left for MaxEnt modeling (Brown et al., 2017).

A total of 19 climatic variables with a spatial resolution in 2.5' for current climate data (average for 1970 - 2000) and four future periods data (BCC-CSM2-MR for 2021-2040, 2041-2060, 2061-2080, 2081-2100) used in the prediction of suitable species distributions were downloaded from WorldClim version 2.1 (https://www.worldclim.org/). We chose two scenarios, namely SSP126,and SSP585, to represent two extreme conditions in the future: a scenario with greenhouse gas well controlled at a low concentration (SSP126, 2.6 W/m2 in 2100) and a scenario with global warming trend without climate policy intervention (SSP585, 8.5 W/m2 in 2100), respectively (Zhao et al., 2021). All the climatic variables were converted to ASCII format using ArcGIS 10.8. To minimize overfitting of the MaxEnt model and ensure prediction accuracy, a Pearson correlation analysis was performed among the 19 climatic variables using the R package *ggpairs* (Emerson et al., 2013), and highly corelated ones were removed (> 0.75, Figure S1) (Dakhil et al., 2019). The correlation of climatic variables was tested all the species together, because one of our targets is to predict the whole genus potential suitable habitat and to raise the protection measures for *Coptis*. Ultimately, by using the software MaxEnt 3.4.1, eleven climatic variables were selected and entered along with the spatial rarefied occurrence data to predict potentially suitable distribution of *Coptis* species.

We set the cross-validation method to randomly selected 75% sites for model training, and the remaining for validation; the model was trained for ten replicate runs (Zhang Y. et al., 2021). Other parameters were kept as the default settings. The calculation result of the MaxEnt model was a grid layer in ASCII format and the value in each grid represented the potential suitable rate of species varying from 0 to 1. Then, we loaded the result into ArcGIS 10.8 to visualize the map of the potential species distribution and regrouped them into four levels as previous studies did: no suitability (0-0.2); low suitability (0.2-0.4); medium suitability (0.4-0.6); and high suitability (0.6-1) (Abolmaali et al., 2018; Zhang et al., 2018; Xu et al., 2019; Kong et al., 2021). Model calibrations and robustness validation were evaluated using the area under the curve (AUC) of the receiver operating characteristics (ROC) curve. The AUC value ranged from 0 to 1 with the following evaluation criteria: poor (0.6–0.7), fair (0.7–0.8), good (0.8–0.9), and excellent (0.9–1) (Zhao et al., 2021). In addition, the contribution rate and jackknife test generated by the MaxEnt model were used to measure the contribution weights of eleven climatic variables. The tool "Distribution changes between binary SDMs" in SDM toolbox v2.5 was used to calculate the changes in distribution area between adjacent time periods.

Result

Features and variation of plastomes genomes

Eleven *Coptis* plastomes were obtained by *de novo* assembly, and deposited in GenBank or National Genomics Data Center (NGDC) with the accession numbers listed in Table S1. The genomes ranged from 153,959 to 154,932 bp in size and contained 113 unique genes (80 protein coding genes, 29 tRNA genes, and four rRNA genes). The whole plastomes had a typical quadripartite structure including a pair of inverted-repeats (IR) regions (26,074 - 26,225 bp), large single copy (LSC) regions (84,112 - 85,178 bp) and small single copy (SSC) regions (17,215 - 17,606 bp). The average GC content ranged from 38.2 to 38.3%.

The results from the mVISTA program analysis showed that most mutation events occurring in the spacer regions and gene of the plastomes were highly conserved, especially in the exon regions (Figure S2). Total sequences had aligned length in 157,727 bp, which had 3,807 variable sites, 1,030 indels and the average nucleotide diversity was 0.0067. Based on the sliding window results, Pi and indels of *Coptis* species were visualized in a Circos map with a window size of each grid of 500 bp (Figure 1A). Most indels in the region were found in ndhF-rpl32. The Pi of a single window varied from 0 to 0.0359 and the indels of a single window varied from 0 to 40. Almost all mutations (92.5%) were located in the SSC and LSC regions, indicating highly conserved IR regions. Moreover, the most hypermutated region was located in the *ycf1* gene. In order to reveal the detailed information of this region, a new sliding-window analysis was performed with settings of 50 bp window length and 25 step size. The most divergent region (Pi > 0.06) of the *ycf1* gene was identified and limited to two tiny regions within 200 bp in size, considered as hotspots (Figure 1B). Compared to the three conventional plastid DNA barcodes (*trnH-psbA*, *rbcL* and *matK*), the two hotspots were nearly twice as high as the second highest barcode *trnH-psbA* and then followed by the complete *ycf1* (Table 1).

Species distribution and diversity pattern

All the 9,182 occurrence records consisted of 15 *Coptis* species that were mainly found in China, Japan, Canada, and United States of America, as well as additionally scattered in the Russian Far East, Korean Peninsula and Greenland (Figure 2A). This clearly indicated that *Coptis* has a typical Eastern Asian and

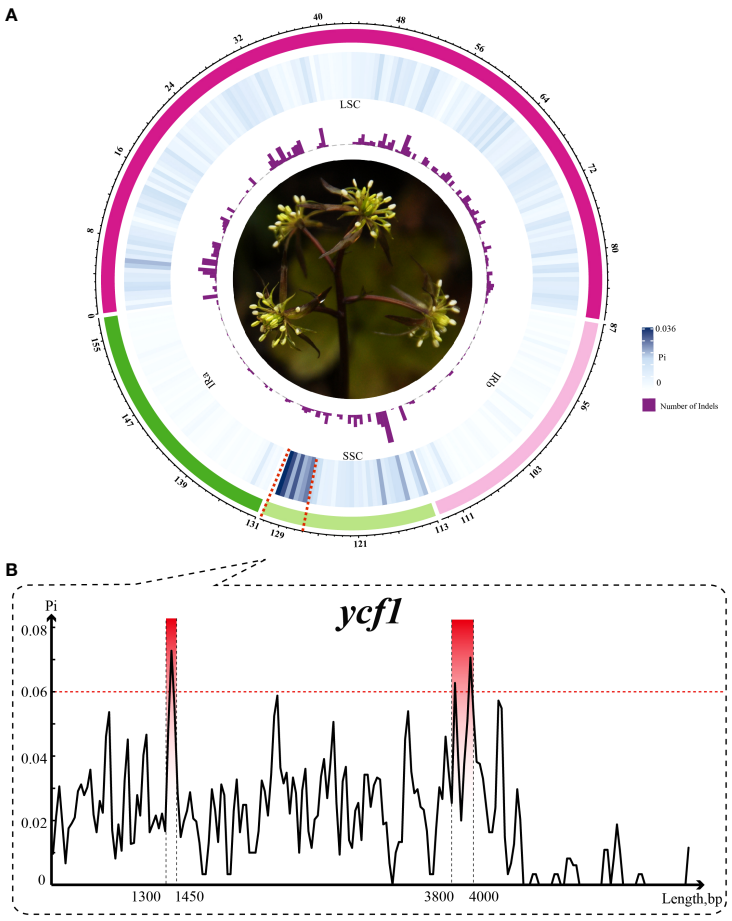


FIGURE 1
Variation of *Coptis* plastomes. **(A)** Circos plot showing the indel and nucleotide diversity of *Coptis*. Circles from the outer to inner area show the following: structure of plastomes indicated by different colors, nucleotide diversity shown by a heatmap and indel count shown by a histogram. Window size of each grid is 500 bp. **(B)** Sliding-window analysis of the *ycf1* gene. The most divergent region (top three region with Pi > 0.06) is indicated.

North American (EA-NA) disjunct distribution pattern. The number of species per grid, a reflection of species diversity, varied from one to six. The grids contained more than three species all located in Eastern Asia, especially in Mid-West China and Japan, which could indicate that these two regions are diversity hotspots of *Coptis*. For North America, the regions along the Pacific Coast Range were slightly more diverse than other regions.

TABLE 1 The variability of the hypervariable markers and universal DNA barcodes.

Markers	Length	Variable sites		information sites		Nucleotide Diversity (Pi)
		Numbers	%	Numbers	%	
rbcL	1428	30	2.10	12	0.84	0.006
matK	1533	95	6.19	39	2.54	0.018
trnH-psbA	316	24	7.59	10	3.16	0.025
rbcL + matK + trnH-psbA	3277	149	4.55	61	1.86	0.012
ycf1	5772	405	7.01	197	3.41	0.020
Two hotspots	417	51	12.23	30	7.19	0.042

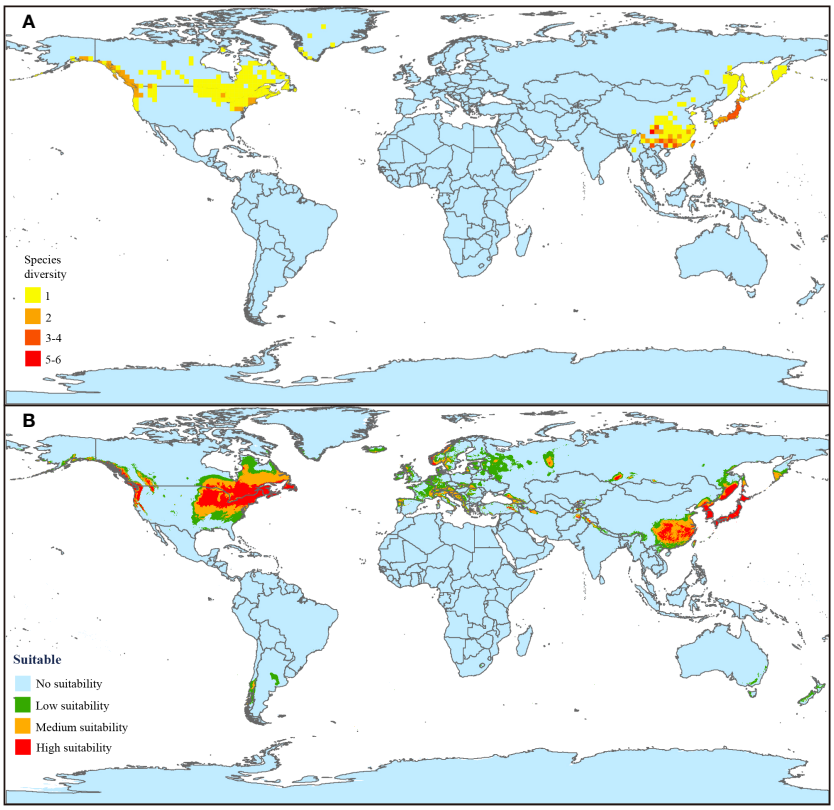


FIGURE 2
Current distribution and suitable habitat prediction of *Coptis*. **(A)** Current species distribution and diversity pattern. The global land area was divided into 2° grid cell by ArcGIS and the number of species per grid were counted to determine the species diversity of *Coptis*. The redder the grid, the higher the species diversity. **(B)** Current potential distribution of *Coptis* under 1970–2000 climate conditions. Four levels of suitability are shown in different colors as follows: no suitability (0-0.2, blue); low suitability (0.2-0.4, green); medium suitability (0.4-0.6, yellow); and high suitability (0.6-1, red).

Phylogeny, divergence time estimation and biogeographical history

The 19 aligned sequences for phylogenetic reconstruction were 191,086 bp in length and 152,783 bp left after trimming by Gblocks. All the *Coptis* species formed a monophyletic group with 100% support. Apparently, *Coptis* species were divided into two large clades, namely Clade I and Clade II (Figure 3 and Figure S3). Clade I (100% bootstrap value) comprised *C. aspleniifolia*, *C. quinquesecta*, *C. japonica*, *C. teeta*, *C. omeiensis*, *C. deltoidei*, *C. chinensis* var. *chinensis* and *C. chinensis* var. *brevisepala*. Species of Clade I were distributed disjunctively among three biogeographical areas (A, C and D). Clade II (100% bootstrap value) comprised *C. trifolia*, *C. quinquefolia*, *C. ramosa* and *C. trifoliolata*, and the latter two formed a monophyly. Additionally, Clade II also distributed disjunctively among three biogeographical areas (B, C and D).

According to the combined dating and RASP results, the *Coptis* genus was most likely to have originated in North America and the Japanese archipelago and the crown age was estimated to be at around 8.49 Mya (95% HPD: 6.56 – 10.23 Mya) in the late Miocene, when the two clades first diverged. In Clade I, *C. trifolia* first split off at around 6.63 Mya when a Bering Land Bridge dispersal (C & D) occurred. Then, the second dispersal (B & C) occurred at around 2.72 Mya when *C. quinquefolia* split off from *C. ramose* and *C. trifoliolata*. Species diversification in Clade II was gradual and successive. *C. aspleniifolia* diverged from Asian species to North American probably around 7.41 Mya. Moreover, *C. quinquesecta* and *C. japonica* successively separated at around 6.21 Mya and 3.73 Mya, corresponding respectively to a dispersal and a vicariance event that occurred between the Chinese mainland and Japanese archipelago with a part of the Russian Far East. This vicariance led to the colonization of *Coptis* in mainland China and the subsequent speciation into five species. In addition, these five species formed a subclade, in which *C. teeta* separated in 2.74 Mya, followed by a divergence of *C. chinensis* from its closest relatives (*C. deltoidei* and *C. omeiensis*) that occurred in 1.87 Mya.

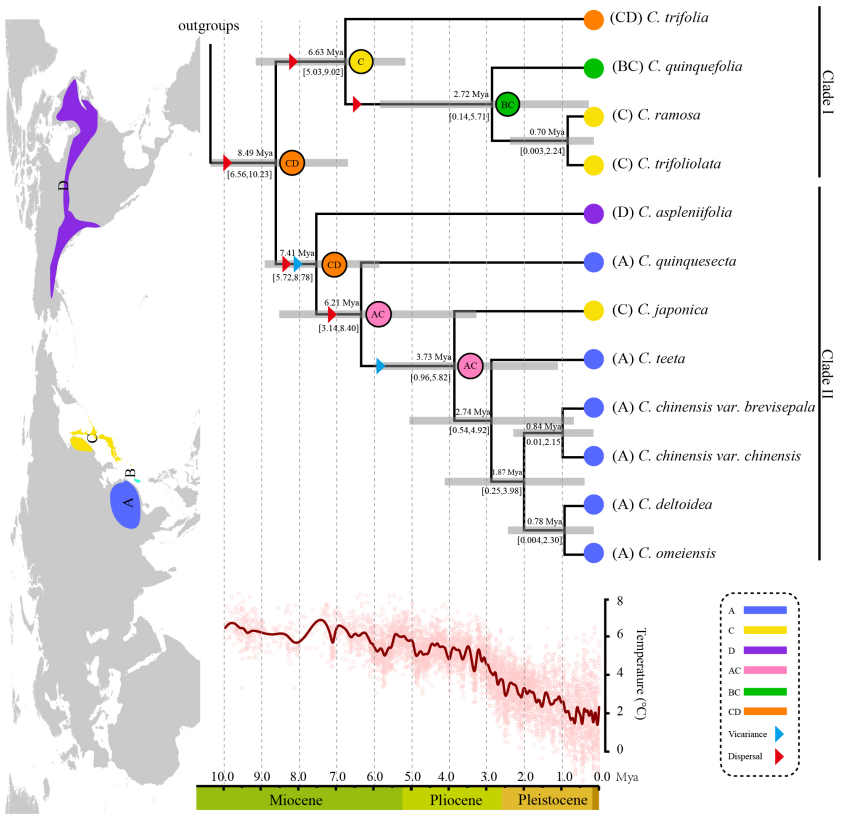


FIGURE 3
Combined dating analyses and ancestral habitat reconstruction of *Coptis*. The global temperature change in the past 10 Ma was obtained from Zachos et al. (Zachos et al., 2008). The insert Northern hemisphere map indicates the species distribution of *Coptis* used in the reconstruction with four defined biogeographical areas as follows: **(A)** mainland China; **(B)**, Taiwan island; **(C)**, Japanese archipelago and a part of the Russian Far East; **(D)**, North America. Dated phylogeny of *Coptis* was derived from Figure S6. Numbers above and under the branches indicate the mean divergence times and 95% confidence interval of each node, respectively. Blue bars indicate the 95% highest posterior density intervals. Letters and colors in the legend represent extant ancestral areas and combination of them. Pie chart labeled with letters at each node indicates the most likely ancestral area.

Modeling validation and dominant climatic variables

The MaxEnt model was used to simulate the suitable habitats of *Coptis*. The final ROC curve (shown in Figure S4), revealed that the average test AUC for the replicate runs was 0.883, with a standard deviation of 0.005, which indicated that the MaxEnt model performed efficiently and reliably in the prediction of suitable habitats.

The eleven selected climatic variables involved in potential suitable habitat prediction are listed in Table 2. The contribution weight of selected climatic variables is shown by the contribution rate and jackknife test result. The variable with the highest contribution rate was precipitation of the driest quarter (Bio 17, contribution rate 38.1%), followed by annual precipitation (Bio 12, contribution rate 37.4%) and annual mean temperature (Bio 1, contribution rate 10.2%). The jackknife test results revealed that the highest gain when used in isolation was still precipitation of the driest quarter (Bio 17, training gain 0.83), followed by annual mean temperature (Bio 1, training gain 0.80) and annual precipitation (Bio 12, training gain 0.79). It is undisputable that these three factors were the most dominant climatic variables in influencing the potential suitable habitat of *Coptis*, while the Bio 1 and Bio12 rankings are different by two assessment methods.

The response curve of the three most dominant climatic variables shown in Figure S5 was generated by the MaxEnt model and illustrates the quantitative relationship between the species presence probability and climatic variables and clearly elucidated the suitability conditions of *Coptis* under different climatic variables. According to the levels of high suitability (> 0.6), the corresponding climatic conditions should be suitable for species growth. In terms of the three dominant climatic variables, the optimum ranges of precipitation of the driest quarter should be more than 25.5 mm, annual precipitation should be more than 844.9 mm and annual mean temperature should range from -3.1 to 19.9°C.

TABLE 2 The description and contribution weight of eleven selected climatic variables.

Variable	Description	Contribution rate	Jackknife test result
Bio1	Annual Mean Temperature	10.2	0.8
Bio2	Mean Diurnal Range (Mean of monthly (Max Temp - Min Temp))	0.2	0.17
Bio3	Isothermality (BIO2/BIO7) (×100)	7.4	0.47
Bio5	Max Temperature of Warmest Month	3.1	0.63
Bio7	Temperature Annual Range (Max Temp of Warmest Month - Min Temp of Coldest Month)	1.9	0.13
Bio8	Mean Temperature of Wettest Quarter	0.2	0.45
Bio9	Mean Temperature of Driest Quarter	0.4	0.63
Bio12	Annual Precipitation	37.4	0.79
Bio15	Precipitation Seasonality (Coefficient of Variation)	0.9	0.48
Bio17	Precipitation of Driest Quarter	38.1	0.83
Bio18	Precipitation of Warmest Quarter	0.2	0.67

Current and future potential habitat prediction and dynamic changes

Based on the current climate condition (1970 - 2000) and occurrence records of *Coptis*, the global potential suitable habitat (suitable rate > 0.2) projected by the MaxEnt model was $1{,}315.0\times10^4$ km^2, which was mostly located in the northern temperate zone, and rarely occurred in tiny regions of New Zealand, Australia, Argentina and Chile (Figure 2B). In China, the potential habitat was 239.6×10^4 km^2, comprising 18% of global suitable area, whose area of the three suitable levels (suitable rate > 0.2) from high to low were 57.7×10^4, 112.2×10^4 and 69.7×10^4 km^2, respectively. The global potential high, medium, and low suitable areas covered 328.9×10^4, 409.9×10^4 and 576.2×10^4 km^2, respectively, and the high suitable area were restricted to the eastern border of the United States of America and Canada, the Pacific Coast Range, most of Northeast Asia, and Mid-West China.

Two climate change scenarios (SSP585 and SSP126) were analyzed to predict the suitable habitat in the future four periods from 2021 to 2100 (Figure 4). As mentioned above, SSP585 is a scenario with no policy intervention and the radiative forcing will rise to 8.5 W/m^2 by 2100. In this scenario, the global potential suitable area (suitable rate > 0.2) has an upward trend followed by a downward trend with the peak occurring at $1{,}648.8\times10^4$ km^2 (2061-2080). Meanwhile, the potential center noticeably shifts north. Correspondingly, the area in China declines continuously from 237.7×10^4 km^2 (2021-2040) to 172.3×10^4 km^2 (2081-2100). In particular, the high suitable area (suitable rate > 0.6) in China decreases most significantly, from 22.7×10^4 km^2 (2021-2040) to merely 1.7×10^4 km^2 (2081-2100). In the SSP126 scenario, the global warming trend may have been suppressed by 2100, the global suitable area shows a slight upward trend (Figure 5). The potential suitable area (suitable rate > 0.2) increases from $1{,}564.8\times10^4$ km^2 (2021-2040) to $1{,}606.4\times10^4$ km^2 (2081-2100) and the center of the suitable area is basically stable. In addition, in contrast to the SSP558 scenario, the suitable area in China shows a gradual upward trend, from 233.0×10^4 km^2 (2021-2040) to 266.1×10^4 km^2 (2081-2100). Also, the decline of the high suitable area (suitable rate > 0.6) is halted and rebounds to 18.6×10^4 km^2 (2081-2100). The shifts of distribution areas between adjacent time periods under two scenarios were vividly presented in Figure S7.

Discussion

Genetic divergence of *Coptis* species and candidate DNA barcodes

As mentioned above, *Coptis* is a genus with enormous value in research and development. Genetic diversity defines the evolutionary potential and resilience of species (Sun et al., 2021). Thus, it is crucially important to evaluate the genetic diversity of this genus. However, there is no comprehensive plastid genetic resource available. In this study, twelve *Coptis* plastomes were found to be highly conserved in genome structure and gene order, and no rearrangement occurred. Some variations were observed in the GC content and genome size. The GC content varied from 38.2 to 38.3% and the genome size varied from 153,959 to 154,932 bp, indicating the existence of genetic diversity. Compared to other medicinal plants, such as *Artemisia* (Pi = 0.0024), *Atractylodes* (Pi = 0.001), *Crataegus* (Pi = 0.00175) and *Ligusticum* (Pi = 0.002), *Coptis* (Pi = 0.0067) is a medicinal genus with relatively high genetic diversity (Kim et al., 2020; Wang Y. et al., 2021; Wu et al., 2022; Wei et al., 2022).

The mVISTA results revealed that, under evolutionary constrains and natural selection pressure, the noncoding region was more variable than the coding region (Zhang H. et al., 2021). According to the sliding-window analysis and Circos map, the LSC and SSC regions were more variable than the IR regions and may be possibly due to copy corrections between IR sequences by gene conversion (Zhao et al., 2018). In

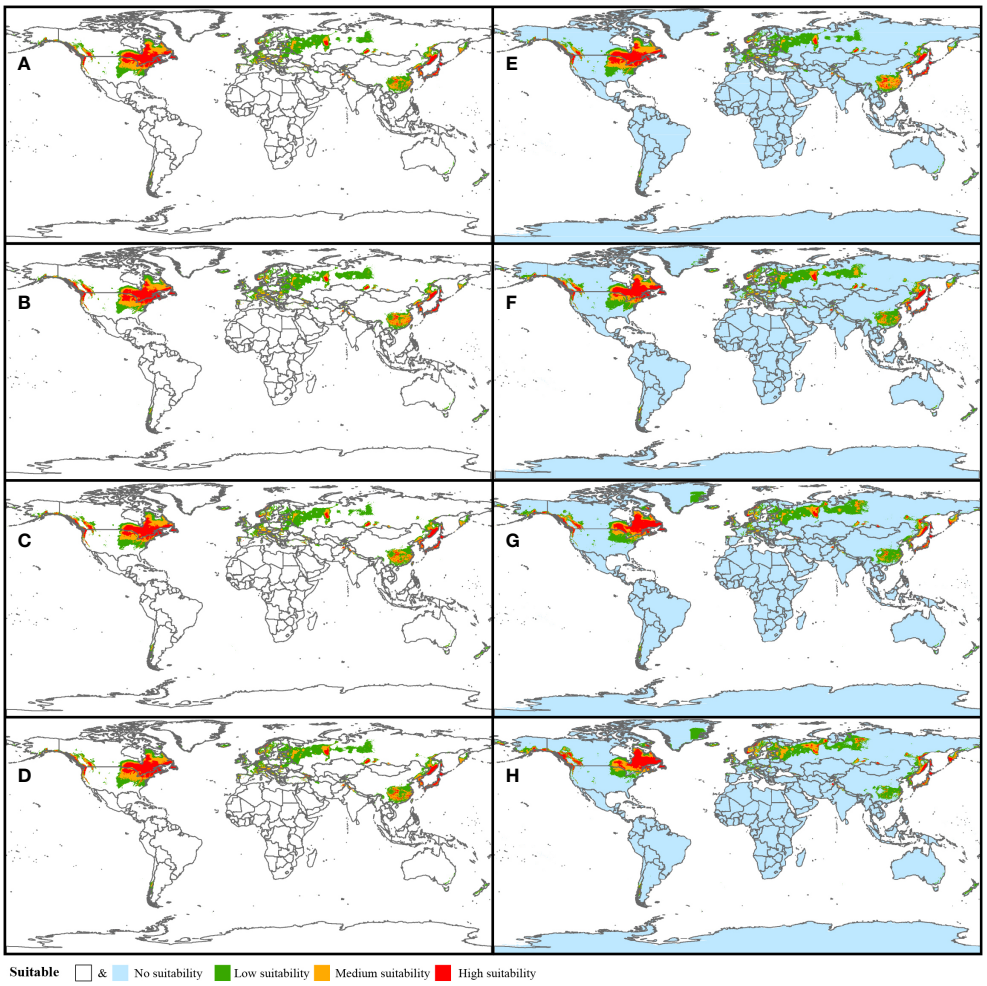


FIGURE 4
Potential habitat prediction of *Coptis* in the future periods: 2021-2040 (SSP126, **A** and SSP 585, **E**), 2041-2060 (SSP126, **B** and SSP 585, **F**), 2061-2080 (SSP126, **C** and SSP 585, **G**) and (SSP126, **D** and SSP 585, **H**).

addition, it also revealed that the mutations were not uniformly or randomly distributed across all regions of the plastome. The aggregation of mutations in certain regions creates hotspots. Hotspots are potential DNA barcode development region that are usually efficient for interspecies discrimination and phylogeny reconstruction (Downie and Jansen, 2015; Song et al., 2017; Pang et al., 2019). Clearly, the hotspot regions of *Coptis* were located in the junction of SSC and IRs, which belonged to the *ycf1* gene. As the second largest gene in the plastid genome, *ycf1* is recognized for its variability in seed plants (Dong et al., 2015). Since it is very long (more than 5 kb) and quite variable in application, two highly mutant regions, namely *ycf1*a (800 bp) and *ycf1*b (1100 bp), of *ycf1* have been developed and used as the most promising plastid DNA barcodes (Dong et al., 2012). Coincidentally, in this study, the precise locations of two hotspots in the *ycf1* were identified, which corresponded to the *ycf1*a and *ycf1*b mutant regions, respectively. Furthermore, the size of these two hotspots were limited to 200 bp. Compared to the three conventional plastid DNA barcodes (*trnH-psbA*, *rbcL* and *matK*), these two regions are much shorter in size but more informative, which qualifies them as mini-barcodes. Notably, mini barcodes may facilitate discrimination of DNA degradation materials, such as herbarium specimens, processed medicinal plant, or even fossil material. Moreover, this strategy of developing taxon-specific barcodes by comparing plastid genome sequences has also been applied in other medicinal plant taxa, including *Panax*, *Senna*, and *Paeonia* (Dong et al., 2014; Yu X. et al., 2020; Yang et al., 2022). Collectively, these two hotspots should be promising mini-barcodes for *Coptis* species identification in future applications.

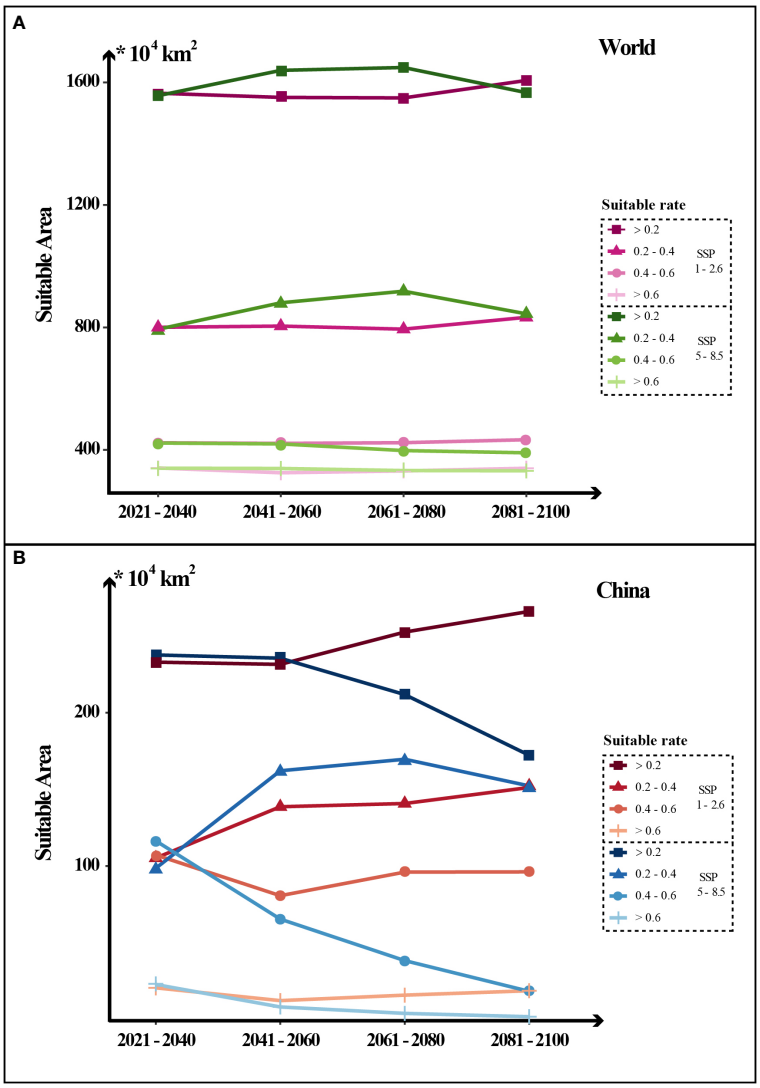


FIGURE 5
Dynamic changes in the area of potential *Coptis* suitable distribution in the world and China in the future four periods under two greenhouse gas emission scenarios (SSP126 and SSP585). The levels of suitability are shown in different line types: low suitability (0.2-0.4); medium suitability (0.4-0.6); high suitability (0.6-1) and total suitability (> 0.2, the sum of high, medium and low levels).

Phylogenetic inferences and evolutionary history

Elucidation of the phylogenetic relationship of *Coptis* is crucial in understanding its evolutionary history. In previous studies, the phylogeny reconstruction of *Coptis* was performed using a small number of DNA loci, such as *trnL-F*, *trnD-T*, *trnH-psbA*, *rpoB*, *accD*, *rbcL*, or limited sampling, which resulted in inconsistent phylogenetic relationships, especially for the four late-diverging species: *C. deltoidea*, *C. omeiensis*, *C. chinensis* var. *chinensis* and *C. chinensis* var. *brevisepala* (He et al., 2014; Xiang et al., 2016; Xiang et al., 2018; Wang et al., 2020). Xiang et al. (2016) suggested that *C. chinensis* var. *brevisepala* formed a clade with *C. deltoidea* and *C. omeiensis*, while *C. chinensis* var. *chinensis* was separated in the early-diverged position based on the matrix combining *trnL-F*, *trnH-psbA* and ITS. Also, Wang et al. indicated that *C. chinensis* var. *brevisepala* was in the early-diverged part of these four species based on the dataset comprising *trnH-psbA*, *rbcL* and *matK*. Importantly, based on adequate sampling and whole plastid genomes with adequate genetic information, this study revealed the most comprehensive phylogeny for *Coptis*. The robust relationship was reconstructed and these controversial branches were fully resolved.

The evolutionary history was clearly shown by estimating the divergence time and biogeography of *Coptis*. The species-level age estimation suggested a crown group age of 8.49 Mya

(95% HPD: 6.56 – 10.23 Mya) in the late Miocene, which was largely consistent with previous results with an age of 9.55 Mya (95% HPD: 6.66–12.92 Mya) determined by DNA barcodes (Xiang et al., 2018). In the middle to late Miocene, the most significant events influencing the evolutionary patterns of global plants were the uplift of the Qinghai-Tibetan Plateau in East Asia and the Rocky Mountains in western North America. These two significant tectonic events in the eastern and western hemispheres had effects on global atmospheric circulation, weathering rates, monsoon and even riverway trend, which might be the main factors underlying the origin and diversification of *Coptis* (Qiu et al., 2011; Pound et al., 2011). *Coptis* probably originated in the Japanese archipelago and a part of the Russian Far East (area C) and North America (area D), and the current distribution of *Coptis* seems to have been shaped by several dispersal and vicariance events. The Bering Land Bridge was the most likely dispersal corridor between the two continents, which has been considered as an explanation for the disjunct distribution of related extant floras in EA and NA. Furthermore, sea level fluctuations and climate cooling since the Pliocene have provided abundant opportunities for promoting dispersal and vicariance among the three Asian biogeographical areas and speciation of *Coptis* occurred rapidly since that time (Haq et al., 1987; Qian and Ricklefs, 2004; Qiu et al., 2011; Xiang, 2020).

It is well accepted that the climatic oscillations, sea-level fluctuations and land bridge configuration promoted speciation and extinction, shaped distribution and diversity pattern of species (Qian and Ricklefs, 2000; Qiu et al., 2009). The diversity center of *Coptis* is located in Japan and the mid-west of mainland China, which are the major parts of the widely recognized biodiversity hotspots "Sino-Japanese Floristic Region" (Hanson et al., 2010). Moreover, for disjunct distributed taxa, diversity is generally higher in East Asia than in North America. The main reason is that East Asia was less influenced by ice sheets during the Quaternary period thus having a lower species extinction rate, and older and more complex topographical features than North America, which may create this biodiversity hotspot in the world with much more species diversity than that in North America (Qian, 2002).

Conservation implications for *Coptis*

Temperature and precipitation are regarded as two of the main variables restricting the range of the majority of terrestrial plant species, as well as that of *Coptis*. As shown by the contribution weight of dominant climatic variables in the results, precipitation is one of the most important factors in future conservation. Understanding the suitable climatic ranges and trend of changes in potential suitable habitat for *Coptis* (precipitation of driest quarter > 25.5 mm, annual precipitation > 844.9 mm and annual mean temperature -3.1 to 19°C), may contribute to provide a basis for ex-situ conservation strategies and the establishment of ex-situ resource nursery for *Coptis* in the future and will also be useful for guiding cultivation and introduction of domestication.

Currently, global climate change greatly affects the distribution of various species and it is also a challenge for all of humanity. Under two greenhouse gas emission scenarios (SSP126 and SSP585), the suitable distribution area of *Coptis* varies largely. In the low-emission condition, the suitable area (suitable rate > 0.2) in world is stable or even shows a slight upward trend, and it is increasing significantly in China. However, without climate policy intervention, the global suitable (suitable rate > 0.2) distribution moves northward and the area in China precipitously declines. The results demonstrate the necessity of tight regulation of greenhouse gas emissions. Confronted by the challenge of climate change, it is important for the state parties and signatories to fulfill the Paris Agreement to reduce their carbon output. These actions should be beneficial not only for the *Coptis* conservation, particularly in China, but also for all lives in the world.

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

YW and JS wrote and revised the manuscript. PQ participated in the experiments. JW, YD, JL, QY, YW, MW and FX collected the materials and analyzed the data. WD, LH and LG conceived and designed the research. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by CACMS innovation Fund (No.CI2021A03909), National Key Research and Development Program of China (2017YFC1703700: 2017YFC1703704), National Natural Science Foundation of China (No.81891014 & No.81874337), Innovation Team and Talents Cultivation Program of National Administration of Traditional Chinese Medicine (No. ZYYCXTD-D-202005) and Genetic Resources Management Project of State Forestry and Grassland Administration (KJZXSA202105).

Acknowledgments

The authors would like to thank professor Shiliang Zhou, Dr. Chao Xu, Kangjia Liu and Enze Li for providing suggestions, and thank the DNA Bank of China in Institute of Botany, Chinese Academy of Sciences for providing materials.

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.2022.1003368/full#supplementary-material

References

Abolmaali, S., Tarkesh, M., and Bashari, H. (2018). MaxEnt modeling for predicting suitable habitats and identifying the effects of climate change on a threatened species, Daphne mucronata, in central Iran. *Ecol. Inform* 43, 116–123. doi: 10.1016/j.ecoinf.2017.10.002

Aguirre Planter, E., Parra Leyva, J. G., Ramirez Barahona, S., Scheinvar, E., Lira Saade, R., and Eguiarte, L. E. (2020). Phylogeography and genetic diversity in a southern north American desert: Agave kerchovei from the tehuacan cuicatlan valley, Mexico. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.00863

Bolger, A. M., Marc, L., and Bjoern, U. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 15), 2114–2120. doi: 10.1093/bioinformatics/btu170

Bouckaert, R., Vaughan, T. G., Barido Sottani, J., Duchene, S., Fourment, M., Gavryushkina, A., et al. (2019). BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. *PloS Comput. Biol.* 15 (4), e1006650. doi: 10.1101/474296

Brown, J., Bennett, J., and French, C. (2017). SDMtoolbox 2.0: the next generation Python-based GIS toolkit for landscape genetic, biogeographic and species distribution model analyses. *PeerJ* 5, e4095. doi: 10.7717/peerj.4095

Castresana, J. (2000). GBLOCLKS: selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17 (4), 540–552. doi: 10.1093/oxfordjournals.molbev.a026334

Chen, D., Pan, Y., Wang, Y., Cui, Y., Zhang, Y., Mo, R., et al. (2021). The chromosome-level reference genome of coptis chinensis provides insights into genomic evolution and berberine biosynthesis. *Hortic. Res.* 8, 121. doi: 10.1038/s41438-021-00559-2

Dakhil, M. A., Xiong, Q., Farahat, E. A., Zhang, L., Pan, K., Pandey, B., et al. (2019). Past and future climatic indicators for distribution patterns and conservation planning of temperate coniferous forests in southwestern China. *Ecol. Indic* 107, 105559. doi: 10.1016/j.ecolind.2019.105559

Dong, W., Li, E., Liu, Y., Xu, C., Wang, Y., Liu, K., et al. (2022). Phylogenomic approaches untangle early divergences and complex diversifications of the olive plant family. *BMC Biol.* 20 (1), 1–25. doi: 10.1186/s12915-022-01297-0

Dong, W., Liu, J., Jing, Y., 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 (4), e35071. doi: 10.1371/journal.pone.0035071

Dong, W., Liu, Y., Li, E., Xu, C., Sun, J., Li, W., et al. (2021a). Phylogenomics and biogeography of catalpa (Bignoniaceae) reveal incomplete lineage sorting and three dispersal events. *Mol. Phylogenet Evol.* 166, 107330. doi: 10.1016/j.ympev.2021.107330

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

Dong, W., Sun, J., Liu, Y., Xu, C., Wang, Y., Suo, Z., et al. (2021b). Phylogenomic relationships and species identification of the olive genus olea (Oleaceae). *J. Syst. Evol.*, 60 (6), 1–18. doi: 10.1111/jse.12802

Dong, W., Xu, C., Li, C., Sun, J., Zuo, Y., Shi, S., et al. (2015). Ycf1, the most promising plastid DNA barcode of land plants. *Sci. Rep.* 5, 8348. doi: 10.1038/srep08348

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 (1), 336–351. doi: 10.1600/036364415X686620

Duan, L., Harris, A., Su, C., Ye, W., Deng, S., Fu, L., et al. (2020). A fossil-calibrated phylogeny reveals the biogeographic history of the cladrastis clade, an amphi-pacific early-branching group in papilionoid legumes. *Mol. Phylogenet Evol.* 143, 106673. doi: 10.1016/j.ympev.2019.106673

Emerson, J. W., Green, W. A., Schloerke, B., Crowley, J., Cook, D., Hofmann, H., et al. (2013). The generalized pairs plot. *J. Comput. Graph Stat.* 22 (1), 79–91. doi: 10.1080/10618600.2012.694762

Feng, Y., Shen, T., Shao, C., Du, H., Ran, J., and Wang, X. (2021). Phylotranscriptomics reveals the complex evolutionary and biogeographic history of the genus tsuga with an East Asian-north American disjunct distribution. *Mol. Phylogenet Evol.* 157, 107066. doi: 10.1016/j.ympev.2020.107066

Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). Circlize implements and enhances circular visualization in r. *Bioinformatics* 30 (19), 2811–2812. doi: 10.1093/bioinformatics/btu393

Hanson, T., Brooks, T. M., Fonseca, G., Hoffmann, M., Lamoreux, J. F., Machlis, G., et al. (2010). Warfare in biodiversity hotspots. *Conserv. Biol.* 23 (3), 578–587. doi: 10.1111/j.1523-1739.2009.01166.x

Haq, B. U., Hardenbol, J., and Vail, P. R. (1987). Chronology of fluctuating sea levels since the Triassic. *Science* 235 (4793), 1156–1167. doi: 10.1126/science.235.4793.1156

He, Y., Hou, P., Fan, G., Arain, S., and Peng, C. (2014). Comprehensive analyses of molecular phylogeny and main alkaloids for coptis (Ranunculaceae) species identification. *Biochem. Syst. Ecol.* 56, 88–94. doi: 10.1016/j.bse.2014.05.002

Huang, X., Ma, L., Chen, C., Zhou, H., Yao, B., and Ma, Z. (2020). Predicting the suitable geographical distribution of sinadoxa corydalifolia under different climate change scenarios in the three-river region using the MaxEnt model. *Plants (Basel)* 9 (8), 1015-1029. doi: 10.3390/plants9081015

Jin, J., Yu, W., Yang, J., Song, Y., dePamphilis, C. W., Yi, T., 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

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 (6), 587–589. doi: 10.1038/nmeth.4285

Kim, G., Lim, C., Kim, J., Kim, K., Lee, J., Yu, H., et al. (2020). Comparative chloroplast genome analysis of artemisia (Asteraceae) in East Asia: insights into evolutionary divergence and phylogenomic implications. *BMC Genomics* 21 (1), 415. doi: 10.1186/s12864-020-06812-7

Kong, F., Tang, L., He, H., Yang, F., Tao, J., and Wang, W. (2021). Assessing the impact of climate change on the distribution of osmanthus fragrans using maxent. *Environ. Sci. pollut. Res. Int.* 28 (26), 34655–34663. doi: 10.1007/s11356-021-13121-3

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

Li, J., Fan, G., and He, Y. (2020). Predicting the current and future distribution of three coptis herbs in China under climate change conditions, using the MaxEnt model and chemical analysis. *Sci. Total Environ.* 698, 134141. doi: 10.1016/j.scitotenv.2019.134141

Liu, Y., Wang, B., Shu, S., Li, Z., Song, C., Liu, D., et al. (2021). Analysis of the coptis chinensis genome reveals the diversification of protoberberine-type alkaloids. *Nat. Commun.* 12 (1), 3276. doi: 10.1038/s41467-021-23611-0

Li, J., Wang, S., Yu, J., Wang, L., and Zhou, S. (2013). A modified CTAB protocol for plant DNA extraction. *Chin. Bull. Bot.* 48 (1), 72–78. doi: 10.3724/SP.J.1259.2013.00072

Ma, K., Wang, X., Feng, S., Xia, X., Zhang, H., Rahaman, A., et al. (2020). From the perspective of traditional Chinese medicine: treatment of mental disorders in COVID-19 survivors. *BioMed. Pharmacother.* 132, 110810. doi: 10.1016/j.biopha.2020.110810

Minami, H., Dubouzet, E., Iwasa, K., and Sato, F. (2007). Functional analysis of norcoclaurine synthase in coptis japonica. *J. Biol. Chem.* 282 (9), 6274–6282. doi: 10.1074/jbc.M608933200

Nge, F. J., Biffin, E., Waycott, M., and Thiele, K. R. (2022). Phylogenomics and continental biogeographic disjunctions: insight from the Australian starflowers (Calytrix). *Am. J. Bot.* 109 (2), 291–308. doi: 10.1002/ajb2.1790

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

Pang, X., Liu, H., Wu, S., Yuan, Y., Li, H., Dong, J., et al. (2019). Species identification of oaks (Quercus l., fagaceae) from gene to genome. *Int. J. Mol. Sci.* 20, (23). doi: 10.3390/ijms20235940

Pound, M. J., Haywood, A. M., Salzmann, U., Riding, J. B., Lunt, D. J., and Hunter, S. J. (2011). A tortonian (late Miocene, 11.61–7.25 ma) global vegetation reconstruction. *Palaeogeogr Palaeocl* 300 (1-4), 29–45. doi: 10.1016/j.palaeo.2010.11.029

Qian, H. (2002). A comparison of the taxonomic richness of temperate plants in East Asia and north America. *Am. J. Bot.* 89 (11), 1818–1825. doi: 10.3732/ajb.89.11.1818

Qian, H., and Ricklefs, R. E. (2000). Large-Scale processes and the Asian bias in species diversity of temperate plants. *Nature* 407 (6801), 180–182. doi: 10.1038/35025052

Qian, H., and Ricklefs, R. E. (2004). Geographical distribution and ecological conservatism of disjunct genera of vascular plants in eastern Asia and eastern north America. *J. Ecol.* 92 (2), 253–265. doi: 10.1111/j.0022-0477.2004.00868.x

Qiu, Y., Fu, C., and Comes, H. (2011). Plant molecular phylogeography in China and adjacent regions: tracing the genetic imprints of quaternary climate and environmental change in the world's most diverse temperate flora. *Mol. Phylogenet Evol.* 59 (1), 225–244. doi: 10.1016/j.ympev.2011.01.012

Qiu, Y., Sun, Y., Zhang, X., Lee, J., Fu, C., and Comes, H. (2009). Molecular phylogeography of East Asian kirengeshoma (Hydrangeaceae) in relation to quaternary climate change and landbridge configurations. *New Phytol.* 183 (2), 480–495. doi: 10.1111/j.1469-8137.2009.02876.x

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Hohna, S., et al. (2012). MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a Large model space. *Syst. Biol.* 61 (3), 539–542. doi: 10.1093/sysbio/sys029

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., Li, Y., Meng, H., Fragniere, Y., Ge, B., Sakio, H., et al. (2020). Phylogeny, taxonomy, and biogeography of pterocarya (Juglandaceae). *Plants (Basel)* 9, (11). doi: 10.3390/plants9111524

Song, Y., Wang, S., Ding, Y., Xu, J., Li, M. F., Zhu, S., et al. (2017). Chloroplast genomic resource of paris for species discrimination. *Sci. Rep.* 7 (1), 3427. doi: 10.1038/s41598-017-02083-7

Sun, J., Wang, Y., Garran, T. A., Qiao, P., Wang, M., Yuan, Q., et al. (2021). Heterogeneous genetic diversity estimation of a promising domestication medicinal motherwort leonurus cardiaca based on chloroplast genome resources. *Front. Genet.* 12. doi: 10.3389/fgene.2021.721022

Tamura, M. (1995). *Phylogeny and classification of the ranunculaceae* (Vienna: Springer Vienna).

Wang, X., Liu, X., Ko, Y., Jin, X., Sun, J., Zhao, Z., et al. (2020). Genetic diversity and phylogeography of the important medical herb, cultivated Huang-lian populations, and the wild relatives coptis species in China. *Front. Genet.* 11. doi: 10.3389/fgene.2020.00708

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 (1), 103. doi: 10.1186/s12864-021-07394-8

Wang, Z., Zhong, C., Li, D., Yan, C., Yao, X., and Li, Z. (2021). Cytotype distribution and chloroplast phylogeography of the actinidia chinensis complex. *BMC Plant Biol.* 21 (1), 325. doi: 10.1186/s12870-021-03099-y

Wei, Z., Xiaoshan, D., Rui, Z., Chunce, G., Lin, L., Guixia, X., et al. (2019). Chloroplast genomic data provide new and robust insights into the phylogeny and evolution of the Ranunculaceae. *Mol. Phylogenet Evol.* 135, 12-21. doi: 10.1016/j.ympev.2019.02.024

Wei, X., Zhang, X., Dong, Y., Cheng, J., Bai, Y., Liu, J., et al. (2022). Molecular structure and phylogenetic analyses of the complete chloroplast genomes of three medicinal plants conioselinum vaginatum, ligusticum sinense, and ligusticum jeholense. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.878263

Wu, X., Luo, D., Zhang, Y., Yang, C., Crabbe., M. J. C., Zhang, T., et al. (2022). Comparative genomic and phylogenetic analysis of chloroplast genomes of hawthorn (Crataegus spp.) in southwest China. *Front. Genet.* 13, 900357. doi: 10.3389/fgene.2022.900357

Xiang, K. (2020). "Phylogeny and diversification of ranunculaceae," in *PhD Dissertation* (University of Chinese Academy of Sciences).

Xiang, K., Erst, A., Xiang, X., Jabbour, F., and Wang, W. (2018). Biogeography of coptis salisb. (Ranunculales, ranunculaceae, coptidoideae), an Eastern Asian and north American genus. *BMC Evol. Biol.* 18 (1), 74. doi: 10.1186/s12862-018-1195-0

Xiang, K., Wu, S., Yu, S., Liu, Y., Jabbour, F., Erst, A., et al. (2016). The first comprehensive phylogeny of coptis (Ranunculaceae) and its implications for character evolution and classification. *PloS One* 11 (4), e0153127. doi: 10.1371/journal.pone.0153127

Xu, X., Liu, L., Cao, X., Long, X., Peng, S., and Zhang, G. (2022). Network pharmacology and molecular docking analysis on molecular targets and mechanism prediction of huanglian jiedu decoction in the treatment of COVID-19. *Digital Chin. Med.* 5 (1), 18–32. doi: 10.1016/j.compbiomed.2022.105389

Xu, D., Zhuo, Z., Wang, R., Ye, M., and Pu, B. (2019). Modeling the distribution of zanthoxylum armatum in China with MaxEnt modeling. *Glob Ecol. Conserv.* 19, e00691. doi: 10.1016/j.gecco.2019.e00691

Yang, Y., Vong, C., Zeng, S., Gao, C., Chen, Z., Fu, C., et al. (2021). Tracking evidences of coptis chinensis for the treatment of inflammatory bowel disease from pharmacological, pharmacokinetic to clinical studies. *J. Ethnopharmacology* 268, 113573. doi: 10.1016/j.jep.2020.113573

Yang, X., Yu, X., Zhang, X., Guo, H., Xing, Z., Xu, L., et al. (2022). Development of mini-barcode based on chloroplast genome and its application in metabarcoding molecular identification of chinese medicinal material radix paeoniae rubra (Chishao). *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.819822

Yan, D., Jin, C., Xiao, X., and Dong, X. (2008). Antimicrobial properties of berberines alkaloids in coptis chinensis franch by microcalorimetry. *J. Biochem.* 70 (6), 845–849. doi: 10.1016/j.jbbm.2007.07.009

Ye, W., Zhu, S., Comes, H., Yang, T., Lian, L., Wang, W., et al. (2022). Phylogenomics and diversification drivers of the Eastern Asian-Eastern north American disjunct podophylloideae. *Mol. Phylogenet Evol.* 169, 107427. doi: 10.1016/j.ympev.2022.107427

Yu, Y., Blair, C., and He, X. (2020). RASP 4: Ancestral state reconstruction tool for multiple genes and characters. *Mol. Biol. Evol.* 37 (2), 604–606. doi: 10.1093/molbev/msz257

Yu, X., Tan, W., Gao, H., Miao, L., and Tian, X. (2020). Development of a specific mini-barcode from plastome and its application for qualitative and quantitative identification of processed herbal products using DNA metabarcoding technique: a case study on senna. *Front. Pharmacol.* 11. doi: 10.3389/fphar.2020.585687

Zachos, J. C., Dickens, G. R., and Zeebe, R. E. (2008). An early Cenozoic perspective on greenhouse warming and carbon-cycle dynamics. *Nature* 451 (7176), 279–283. doi: 10.1038/nature06588

Zeng, L., Qiang, Z., Sun, R., Kong, H., Ning, Z., and Hong, M. (2014). Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. *Nat. Commun.* 5, 4956. doi: 10.1038/ncomms5956

Zhang, D., Gao, F., Jakovlic, I., Zou, H., Zhang, J., Li, W., 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, X., Landis, J., Wang, H., Zhu, Z., and Wang, H. (2021). Comparative analysis of chloroplast genome structure and molecular dating in myrtales. *BMC Plant Biol.* 21 (1), 219. doi: 10.1186/s12870-021-02985-9

Zhang, Y., Tang, J., Ren, G., Zhao, K., and Wang, X. (2021). Global potential distribution prediction of xanthium italicum based on maxent model. *Sci. Rep.* 11 (1), 16545. doi: 10.1038/s41598-021-96041-z

Zhang, K., Yao, L., Meng, J., and Tao, J. (2018). Maxent modeling for predicting the potential geographical distribution of two peony species under climate change. *Sci. Total Environ.* 634, 1326–1334. doi: 10.1016/j.scitotenv.2018.04.112

Zhao, Y., Deng, X., Xiang, W., Chen, L., and Ouyang, S. (2021). Predicting potential suitable habitats of Chinese fir under current and future climatic scenarios based on maxent model. *Ecol. Indic* 64, 101393. doi: 10.1016/j.ecoinf.2021.101393

Zhao, Z., Wang, X., Yu, Y., Yuan, S., Jiang, D., Zhang, Y., et al. (2018). Complete chloroplast genome sequences of dioscorea: Characterization, genomic resources, and phylogenetic analyses. *Peerj* 6, 1-13. doi: 10.7717/peerj.6032

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 24 November 2022
DOI 10.3389/fpls.2022.1047592

OPEN ACCESS

EDITED BY
Wenpan Dong,
Beijing Forestry University, China

REVIEWED BY
Xiwen Li,
Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, China
Lihong Xiao,
Zhejiang Agriculture and Forestry University, China

*CORRESPONDENCE
Ningjia He
hejia@swu.edu.cn

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

RECEIVED 18 September 2022
ACCEPTED 24 October 2022
PUBLISHED 24 November 2022

CITATION
Zeng Q, Chen M, Wang S, Xu X, Li T, Xiang Z and He N (2022) Comparative and phylogenetic analyses of the chloroplast genome reveal the taxonomy of the *Morus* genus. *Front. Plant Sci.* 13:1047592. doi: 10.3389/fpls.2022.1047592

Comparative and phylogenetic analyses of the chloroplast genome reveal the taxonomy of the *Morus* genus

Qiwei Zeng, Miao Chen, Shouchang Wang, Xiaoxiang Xu, Tian Li, Zhonghuai Xiang and Ningjia He*

State Key Laboratory of Silkworm Genome Biology, Southwest University, Chongqing, China

Mulberry (genus *Morus*) is an economically important woody plant with an altered ploidy level. The variable number of *Morus* species recognized by different studies indicates that the genus is in need of revision. In this study, the chloroplast (CP) genomes of 123 *Morus* varieties were *de novo* assembled and systematically analyzed. The 123 varieties represented six *Morus* species, namely, *Morus alba*, *Morus nigra*, *Morus notabilis*, *Morus rubra*, *Morus celtidifolia*, and *Morus serrata*. The *Morus* CP genome was found to be 158,969~159,548 bp in size with 125 genes, including 81 protein coding, 36 tRNA, and 8 rRNA genes. The 87 out of 123 mulberry accessions were assigned to 14 diverse groups with identical CP genome, which indicated that they are maternally inherited and share 14 common ancestors. Then 50 diverse CP genomes occurred in 123 mulberry accessions for further study. The CP genomes of the *Morus* genus with a quadripartite structure have two inverted repeat (IR) regions (25,654~25,702 bp) dividing the circular genome into a large single-copy (LSC) region (87,873~88,243 bp) and small single-copy (SSC) region (19,740~19,994 bp). Analysis of the phylogenetic tree constructed using the complete CP genome sequences of *Morus* revealed a monophyletic genus and that *M. alba* consisted of two clades, *M. alba* var. *alba* and *M. alba* var. *multicaulis*. The Japanese cultivated germplasms were derived from *M. alba* var. *multicaulis*. We propose that the *Morus* genus be classified into six species, *M. nigra*, *M. notabilis*, *M. serrata*, *M. celtidifolia*, *M. rubra*, and *M. alba* with two subspecies, *M. alba* var. *alba* and *M. alba* var. *multicaulis*. Our findings provide a valuable resource for the classification, domestication, and breeding improvement of mulberry.

KEYWORDS

Mulberry, Chloroplast genome, Phylogenetic tree, Taxonomy, Morus alba

Introduction

Mulberry (*Morus* L., Moraceae) (Group, 2009) comprises a variable number of species, with the first 7 described by Linnaeus (1753). The traditional taxonomy of *Morus* is often based on minor morphological differences (Gardner et al., 2021). Various researchers identified 5, 8, 13, 16, 24, and 35 *Morus* species using morphological and/or molecular methods (Bureau, 1873; Koidzumi, 1930; Hotta, 1954; Zhou and Gilbert, 2003; Zeng et al., 2015; Jain et al., 2022). The classification of the *Morus* genus based on morphology did not truly reflect the phylogenetic relationships (Jiao et al., 2020). In 2015, we proposed eight species in the *Morus* genus on the basis of ITS (internal transcribed spacer) sequences, which were recently verified by an analysis of population genetics (Jiao et al., 2020). A recent investigation of Moraceae revealed that the genus *Morus* is a monophyletic group without *M. mesozygia* and *M. insignis*, and the delimitation of *M. alba* may be worth further investigating (Gardner et al., 2021). However, the genome-based taxonomy of the genus *Morus* remains unexplored (Jiao et al., 2020).

Since the first mulberry genome (*M. notabilis*) was published (He et al., 2013), six other mulberry genomes, including chromosome-level genomes, have been reported (Jiao et al., 2020; Muhonja et al., 2020; Jain et al., 2022; Xia et al., 2022; Xuan et al., 2022), which represent excellent reference genomes for genomic and population analyses of mulberry resources. Different chromosome numbers (14, 28, 35, 42, 49, 56, 84, 112, 126, and 308) with various ploidy levels have been reported in mulberry (Tikader and Kamble, 2008; Xuan et al., 2022). For example, black mulberry (*M. nigra*) is a polyploid with 308 chromosomes (Agaev and Fedorova, 1970; SB et al., 1990), and *Morus serrata* is a natural polyploid with 56 or 84 chromosomes (SB and Rajan, 1989). Variable ploidy levels are often observed in white mulberry (*M. alba*) (Xuan et al., 2019). The variable ploidy levels in *Morus* adversely affect population genetic analyses for taxonomic purposes. Mulberry has been cultivated by farmers for over 5000 years (He et al., 2013), and many varieties or cultivars have been generated by natural and artificial breeding selection. More than 2600, 1500, and 1120 mulberry germplasm resources were recorded in China, Japan, and India, respectively (Vijayan et al., 2011; Vijayan and B.da Silva, 2011). The evolutionary relationships of these cultivars or varieties remain unclear.

Chloroplast, a plant cell organelle with its own genome, is essential for the growth and development of plants. Compared with the large nuclear genome, chloroplast genomes are smaller. CP genomes with their numerous advantages for plant phylogeny reconstruction, including a relatively conserved rate of evolution and usually uniparental inheritance, provide an important resource for elucidating morphological evolution (Gitzendanner et al., 2018; Li et al., 2021; Hua et al., 2022). CP genomes also provide critical insights into historically difficult relationships of the major angiosperm subclades (Moore et al., 2007; Moore et al., 2010; Stull et al., 2015; Sun et al., 2016; Li H-T. et al., 2019). We proposed that the CP genomes possibly provide insights into the evolution and taxonomy of the *Morus* genus. Since the CP genome of *M. indica* var. K2 was first obtained (Ravi et al., 2006), those of *M. mongolica* (Kong and Yang, 2016), *M. alba* var. *atropurpurea* (Li et al., 2016), *M. notabilis* (Chen et al., 2016), *M. alba* var. *multicaulis* and *M. cathayana* (Kong and Yang, 2017), *M. alba* (Luo et al., 2019; He et al., 2020), etc., have been reported. However, a large-scale comparative analysis of the CP genome across the *Morus* genus has not yet been conducted.

Therefore, the purpose of this study is to conduct a large-scale comparative genomic analysis of *Morus* CP genomes and reconstruct phylogenetic tree based on CP genomes to explore the taxonomy of genus *Morus*. The evolutionary relationships of mulberry accessions were also explored based on their CP genomes. These results provide an important information for the classification, domestication, and breeding improvement of mulberry.

Material and methods

Sample collection and sequencing

Morus serrata was collected from Jilong, Tibet Autonomous Region, China, and propagated at the Mulberry Germplasm Nursery at Southwest University. *Morus celtidifolia* was identified by Professor Elizabeth Makings from Arizona State University, USA. *Morus notabilis* was collected from a pristine forest in Ya'an, Sichuan Province, China. *Morus yunnanensis* was obtained from the Institute of Sericulture and Apiculture, Yunnan Academy of Agricultural Sciences, Mengzi, Yunnan Province, China. *Morus nigra* was collected from Yutian County, Xinjiang Uygur Autonomous Region, China. Other samples were obtained from the Mulberry Germplasm Nursery at Southwest University, China. For each sample, 10 μg genomic DNA was extracted from young leaves according to a standard cetyltrimethylammonium bromide protocol for the subsequent construction of sequencing libraries. Specifically, sequencing libraries with an average insert size of 350 bp were constructed according to the Illumina standard protocol, after which they were sequenced by BGI-Shenzhen (Shenzhen, China) using the Illumina HiSeq XTen or MGISEQ-2000 platform (Illumina, San Diego, CA, United States) to generate 150-bp paired-end reads. The raw data of 35 samples have been deposited in the CNGB Sequence Archive of the China National GeneBank Database (CNGBdb) under accession number CNP0001407. Using these data and publicly available genomic data downloaded from the NCBI or CNGB database, the *Morus* CP genomes were studied (Supplementary 1). The adapters and low-quality sequences

were removed using the program fastp (Chen et al., 2018) from the raw reads to obtain clean reads for the subsequent analyses.

De novo assembly and annotation of the chloroplast genome

NOVOPLasty (version 4.3) (Dierckxsens et al., 2017) and GetOrganelle (v1.7.6.1) (Jin et al., 2020), which were developed for the *de novo* assembly of organelle genomes, were used for assembling CP genomes. For NOVOPLasty, default parameters were applied, with the following exceptions: read length (100 or 150), genome range (150,000–170,000), and K-mer optimized. For GetOrganelle, default parameters were applied, with the following exception: the heyebai chloroplast genome sequence (KU981119) as a reference sequence. Because the two haplotypes are present in the same proportion in a cell (Wang and Lanfear, 2019), we then selected the haplotype with the same SSC orientation as that in the CP genome sequences for further analyses.

The complete CP genomes were annotated in CPGAVAS2 (Shi et al., 2019) with default parameters. The ambiguous gene positions were manually corrected by NCBI BLASTN searches. All transfer RNA genes were confirmed on the tRNAscan-SE 2.0 web server with default settings (Lowe and Chan, 2016). Their high-quality graphical maps were drawn by OGDRAW (Lohse et al., 2013) with default parameters. All annotated chloroplast genome sequences were submitted to GenBank through BankIt (https://www.ncbi.nlm.nih.gov/WebSub/index.cgi).

Comparative chloroplast genome analysis

The mVISTA program (http://genome.lbl.gov/vista/mvista/about.shtml) was employed to determine the differences in the whole chloroplast genomes of *M. notabilis* (MK211167), *M. serrata* (MT154044), *M. celtidifolia* (MT154045), *M. alba* var. *multicaulis* (OP153908), *M. alba* var. *alba* (OP153917), *M. nigra* (OP153918), *M. alba* var. *indica* (OP153922), and *M. rubra* (OP161259), in the Shuffle-LAGAN mode with *M. notabilis* as the reference genome.

Sequence divergence analysis

MAFFT v7.455 software (Katoh and Standley, 2013) was employed to align the CP genomes of 50 *Morus* accessions. DnaSP v5.10 software (Librado and Rozas, 2009) was used to identify rapidly evolving molecular markers with a sliding window analysis (window length and step size set as 500 and 250 bp, respectively). The R package ggmsa (Zhou et al., 2022) was used to visualize multiple sequence alignments of target regions from 50 CP genomes.

Phylogenetic analysis

We assigned mulberry accessions with the identical CP genomes to a group, therefore 87 out of 123 mulberry accessions were assigned to 14 different groups and other 36 accessions were classified to an unclassified group. In 14 different groups, each group randomly selected a representative CP genome, together with other 36 CP genomes from the unclassified group, to form 50 CP genomes for phylogenetic analysis. The CP genome sequences of *Broussonetia papyrifera* (*Broussonetia* genus, GenBank: MZ662865), *Ficus carica* (*Ficus* genus, GenBank: KY635880) and *Morus mesozygia* (*Afromorus* genus, GenBank: MZ274134) (Gardner et al., 2021) were selected as outgroups. All complete CP genome sequences were aligned using MAFFT v7.455 (Katoh and Standley, 2013), and the alignments were trimmed with trimAl v1.4.rev15 (Capella-Gutiérrez et al., 2009). IQ-TREE 2.0 (Minh et al., 2020) was employed to construct a maximum likelihood (ML) phylogenetic tree with 1000 bootstrap replicates. Finally, the phylogenetic tree was edited using iTOL 5.0 (https://itol.embl.de/) (Letunic and Bork, 2019) and Adobe Photoshop® CC (Adobe Systems Inc., California, U.S.A.).

Results

Features of the *Morus* species chloroplast (CP) genomes

We *de novo* assembled 123 complete CP genomes of *Morus* species with sizes ranging from 158,969 to 159,548 bp (Figure 1; Supplementary 1). These CP genomes display a typical circular quadripartite architecture, with an LSC region (87,873~88,243 bp) and an SSC region (19,740~19,994 bp) separated by two inverted repeat (IR) regions (25,654~25,702 bp) (Supplementary 1). All CP genomes showed similar total GC contents (ranging from 36.13% to 36.21%). The largest length change in CP genome sequences occurred upstream of psbA of *M. alba* var. *indica* with a 135 bp missing sequence, which led *M. alba* var. *indica* to have the shortest CP genome (Supplementary 1).

The mulberry accessions with same CP genomes were assigned to a group, 87 out of 123 accessions were assigned to 14 diverse groups and other 36 accessions with different CP genomes were assigned to an unclassified group (Table 1 and Supplementary 1), which indicated that these 87 mulberry accessions were inherited maternally and shared 14 common ancestors. Of them, the largest group, ONE, comprises 28 *Morus* accessions, including 9 Japanese mulberry accessions and 9

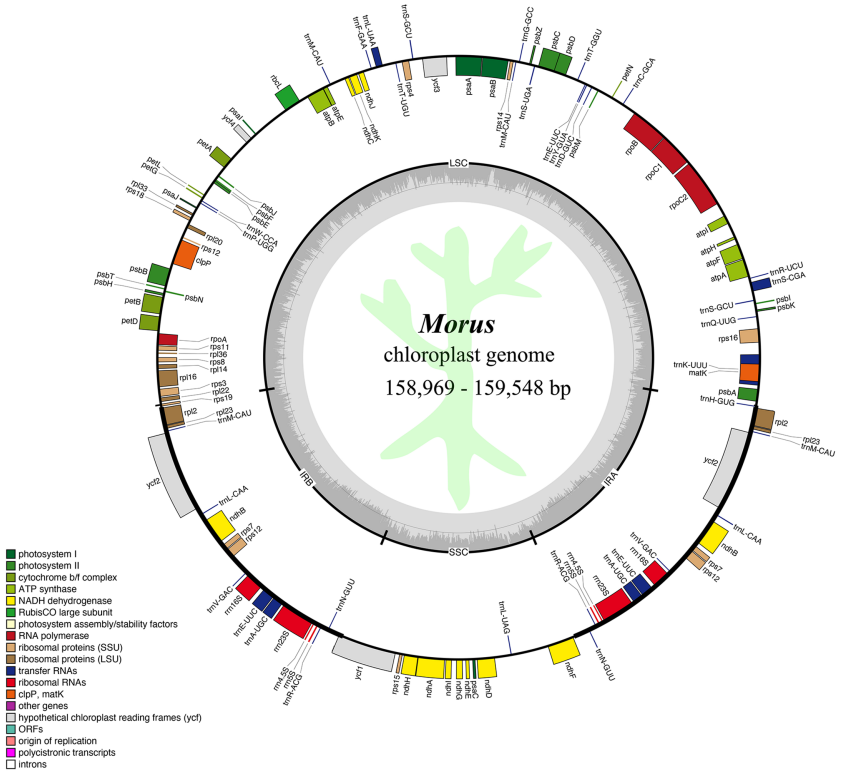


FIGURE 1
Genome map of the average *Morus* CP genome studied in this work. The inner circle represents the quadripartite structure, with an LSC, an SSC, and inverted repeat regions (IRA and IRB) and GC content shown in dark gray and AT content shown in light gray. The external circle represents gene content, with those outside the circle transcribed counterclockwise, while those located inside are transcribed clockwise. Genes are colored according to functional groups defined in the legend shown in the bottom left. The script in light green represents Jiaguwen mulberry.

Husang (*M. alba var. multicaulis*) accessions (Table 1). The nine Husang germplasms include heyebai (Husang-32), Jiantouheyebai, Husang-103, Husang-26, Husang-37, Husang-10, and Husang-60. Group TWO consists of 13 *Morus* accessions, including 12 Indian mulberry samples. Group THREE consists of 10 different Chuansang (*M. notabilis*) samples collected in a pristine forest around Ya'an, Sichuan Province. Group SIX comprises 4 samples, including one maternal parent and its three hybrid progenies. Group EIGHT comprises 4 Japanese samples, Yamasou-159, Ichinose, Fengwei-Ichinose, and Shinichinose. In 14 different groups, each group randomly selected a representative CP genome, together with other 36 CP genomes from the unclassified group Table 1, to form 50 diverse CP genomes occurred in 123 mulberry accessions for further analysis. And these 50 CP genomes were deposited in NCBI with GenBank accession numbers OP153890-OP153924, OP161257-OP161267, OP380682-OP380686, MK211167, MT154044, MT154045, and OP142713 for further analysis (Supplementary 1).

We annotated 125 genes in each *Morus* CP genome, consisting of 81 protein-coding genes, 8 rRNAs, and 36 tRNAs (Figure 1). In detail, 17 duplicated genes in the IR region were identified, including 6 protein-coding genes, 4 rRNA genes, and 7 tRNA genes. Ten protein-coding genes and 1 tRNA gene exist in the SSC region, while 59 protein-coding genes and 21 tRNA genes are present in the LSC region. All annotated genes were concatenated into a supermatrix, whose sizes in the 50 CP genomes ranged from 109,073 to 109,159 bp (Supplementary 1). Twenty-six out of 50 supermatrices showed identical sequences in ten separate groups (Supplementary 2). Of them, the largest three groups contained five, four, and three members, respectively, indicating high conservation.

There were 22 genes containing introns in all *Morus* CP genomes (Supplementary 3). Among these genes, 7 are tRNA genes, and 15 are protein-coding genes. Most genes have only a single intron, whereas the clpP and ycf3 genes contain two introns (Supplementary 3). rps12 is a trans-splicing gene composed of three exons, containing one 5' exon located in the LSC region and two 3' exons located in the IR region (Figure 1). Compared to other intron-containing genes, the trnK-UUU gene embodied the matK genes and has the largest intron (2,553-2,563 bp) (Supplementary 3).

TABLE 1 Group information of the 123 mulberry germplasm accessions.

Groups	Sample names of mulberry germplasm accessions
ONE	Heyebai, Jiantouheyebai, Husang-103, Heyebai-reseq, Husang-26, Husang-37, **Nezumigaeshi**-2, **Kenmochi**, Tuosang-25, **Kumonryu**, **Kenmochi-2**, **Gunmaakagi**, **Shuangtou-Kenmochi**, Tieyezi, Liyeda, Husang-10, Husang-60, zhangyasang, Heyebai, **Kairyouwasejumonji**, 6071, Chaoxiansang, Guizhoumaosang-31, **Kenmochi**, Gelujiya, Huosang-1, NXS, **Dateakagi**
TWO	kanva2, SL2, Mulberry-MJ049, Mulberry-MJ036, Mulberry-MJ038, Mulberry-MJ033, Mulberry-MJ041, Mulberry-MJ039, Mulberry-MJ047, Mulberry-MJ035, kanva2, Mulberry-MJ042, Mulberry-MJ032
THREE	*M. notabilis*_2, *M. notabilis*_4, *M. notabilis*_5, *M. notabilis*_reseq, wild_mulberry_S16, f_CHS_1, m_CHS_1, m_CHS_3, f_CHS_2, m_CHS_2
FOUR	Shidian-6, Dazhongsang, Linxianlusang, RL0424, Shaqi-1, Huanglutou
FIVE	Huasang, Cangxisang, Taiwanchangguosang, SL1, Mulberry-MJ044, Mulberry-MJ045
SIX	Hybrid-13, Hybrid-n13, 12, Lunjiao109x12
SEVEN	Chuizhisang, Chuizhisang, Chuizhisang, ZCS
EIGHT	**Yamasou-159**, **Ichinose**, **Fengwei-Ichinose**, **Shinichinose**
NINE	*Morus multicaulis*, **Shaansang305**
TEN	Huanggelu, Heigelu
ELEVEN	*Morus yunnanensis*, *Morus yunnanensis*
TWELVE	Lunjiao109_G, Lunjiao109_C
THIRTEEN	Xiongyue-107, Jizhuasang
FOURTEEN	T1, Tseed
UNCLASSIFIED	Pisang-2, Yun-7, *Morus macroura*, *Morus australis*, *Morus wittiorum*, Yun-6, **Kairyonezumigaeshi**, Liquan-1, Qinbasang, JiLongSang, *Morus nigra*, *Morus celtidifolia*, Zhaisang-1, Jiaqing-9, Guapiaosang, Dabaie, Shuangcheng, Heilujiesang, Jainzhizi, **Yamasou -106**, Basailuona, Chenkou-3, JPZ, Tianquan-12, *Morus rubra*, *Morus cathayana*, Gui-23, Kangqin283, Hongye, Baiyuwang, Zhenzhubai, Shuisang, WG120_1, Husang-192, Pisang, Luohang-5

Sample names in bold are from Japanese mulberry germplasm accessions. Lunjiao109_G and Lunjiao109_C represented samples collected in Guangdong province and Chongqing province, China, respectively.

Similarity analysis and nucleotide diversity of CP genomes

The sequence homology of the *Morus* species was investigated with *M. notabilis* as a reference using mVISTA software (Figure 2). The nucleotide variability (Pi) was calculated to further confirm the sequence variations (Figure 3). The *Morus* CP genomes were highly conserved and displayed similar structures and gene orders (Figure 2). The divergence level of the noncoding regions was higher than that of the coding regions. The protein-coding regions were highly conserved, and the ndhF genes displayed obvious polymorphism (Figure 2). The Pi values were rather low, ranging from 0 to 0.00442 among the 50 CP genomes, and 2 hotspot regions were identified with Pi >0.003 (rps16-trnQ-UUG and trnL-UAG-ndhF) (Figure 3). No highly variable loci were detected in the IR regions, and the nucleotide diversity values were significantly lower than those in the single-copy regions (Figure 3). Because of the highly conserved sequences, structure, and size of the CP genomes of *Morus*, no obvious hypervariable regions were noted (Figures 2, 3). As a result, the complete CP genomes were considered to distinguish *Morus* species.

Phylogenetic analysis

In this study, the 50 CP genomes representing 123 mulberry samples were utilized to explore the phylogenetic positions of *Morus* species. Because of the highly conserved coding-region sequences in the *Morus* CP genomes, the complete genomes were used to construct the maximum-likelihood (ML) tree. As illustrated in Figure 4, the phylogenetic tree was divided into five clades. Among them, Outgroup is a clade containing three different genera (*Ficus*, *Broussonetia*, and *Afromorus*) at the root. *Ficus carica* was clustered with *Broussonetia papyrifera*, which was a sister genus of *Morus*, indicating a close relationship between *Ficus* and *Broussonetia*. *Morus mesozygia*, an outgroup member, was recognized as a *Morus* species native to Africa and belongs to the *Afromorus* genus. *M. celtidifolia*, a species native to America, was an independent clade. *M. notabilis*, native to Sichuan Province, and *M. yunnanensis*, native to Yunnan Province, formed a clade. Black mulberry (*M. nigra*) was a clade. White mulberry (*M. alba*) was the most complex and largest clade and was further divided into two subclades, *M. alba* var. *alba* and *M. alba* var. multicaulis, indicating that there were two subspecies of *M. alba* species. The *M. alba* var. *multicaulis* subclade comprised three subgroups containing all Husang, *M. alba* var. *indica*, and Japanese mulberry accessions. The *M. alba* var. *alba* subclade contained three subgroups, including the red mulberry (*M. rubra*), *M. serrata*, and a wild mulberry collected in Tibet, China. In addition, the mulberry resources (*M. alba* var. Taiwanchangguosang, *M. alba* var. *shuisang*, *M. wittiorum*, *M. alba* var. Yun7 and *M. alba* var. Yun6) with long fruits (over 4 cm) were clustered in the *M. alba* var. *alba* subclade. At the same time, two *M. alba* var. *atropurpurea* germplasms (*M. alba*

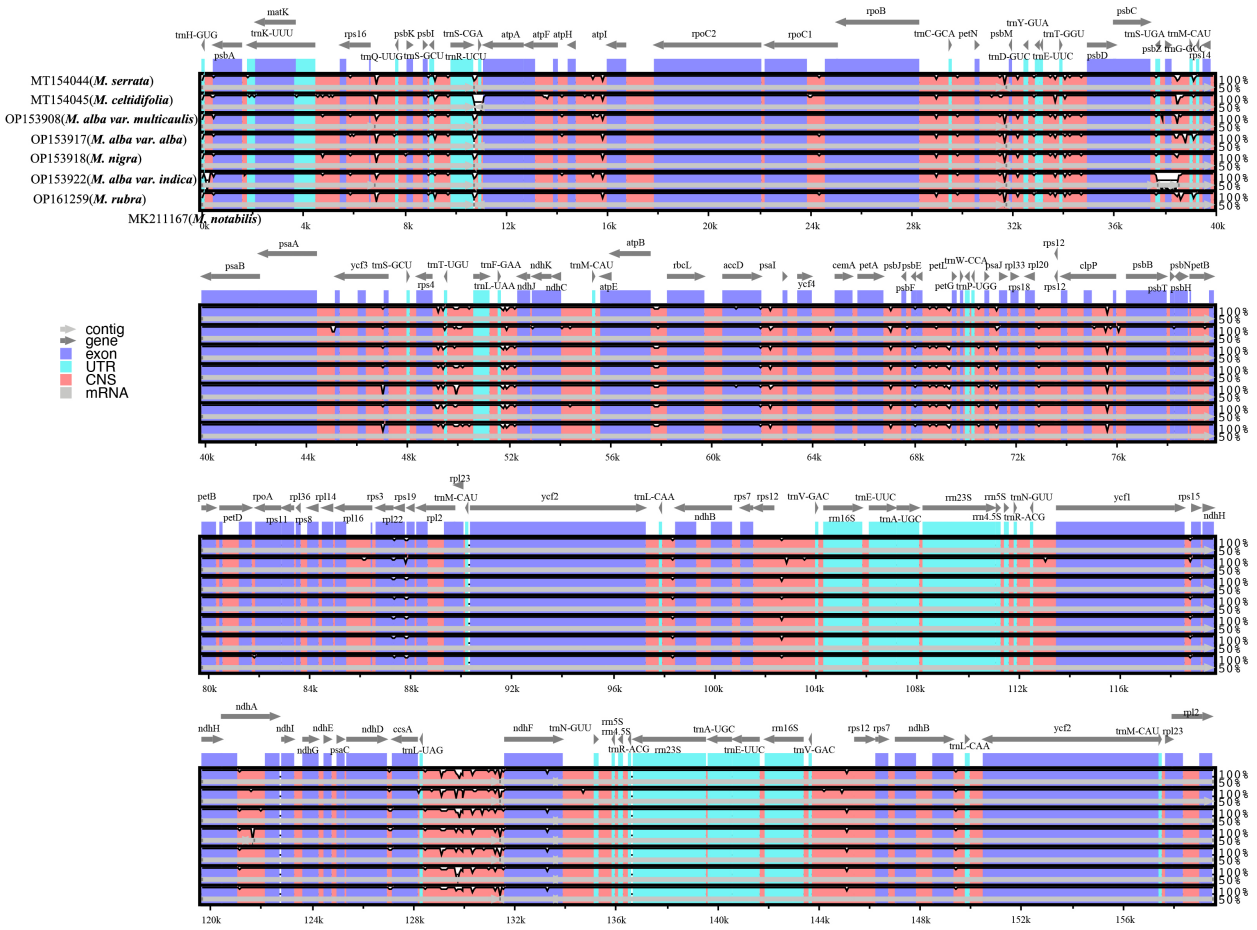


FIGURE 2
Sequence identity plot comparing the chloroplast genomes of six *Morus* species with *M. notabilis* as a reference. The vertical scale indicates the percentage of identity, ranging from 50 to 100%. The horizontal axis indicates the coordinates within the CP genome.

var. Lunjiao109 and *M. alba* var. Kanqin283) were placed in the *M. alba* var. *alba* subclade.

Discussion

Morus CP genome characterization

Maternally inherited CP genomes provide useful information for phylogenetic reconstruction (Bruun-Lund et al., 2017; Gitzendanner et al., 2018; Li et al., 2021; Wang et al., 2021; Zhang et al., 2022). Although some *Morus* CP genomes (Chen et al., 2016; Kong and Yang, 2016; Li et al., 2016; Kong and Yang, 2017; Luo et al., 2019; He et al., 2020) have previously been reported since the first was reported in 2006 (Ravi et al., 2006), there is a lack of large-scale comparative analysis of these genomes. Some *Morus* CP genomes deposited in the NCBI database were reference-based assemblies (Chen et al., 2016; Kong and Yang, 2016; Li et al., 2016), which may lack some useful information. For example, the CP genome of *M. notabilis* showed length differences between *de novo* assembly (GenBank: MK211167, size: 159,548 bp) and reference-based assembly (GenBank: KP939360, size: 158,680 bp). In addition, two indels over 40 bp were detected in the CP genome of the *de novo* assembly (GenBank: OP153912, size: 159,200 bp) compared with the reference-based assembly (GenBank: KU981119, size: 159,103 bp) using the same raw data. The performance of the reference-based assembly was dependent on the references employed (Scheunert et al., 2020). As a result, all *Morus* CP genomes were *de novo* assembled in this study, and *Morus* CP genomes of the reference-based assembly were not included. The size of *Morus* CP genomes ranged from 158,969 to 159,548 bp, which was larger than the first *Morus* CP genome (158,484 bp) (Ravi et al., 2006) and suggested that CP genome length in *Morus* was highly conserved. GC content is often considered an important

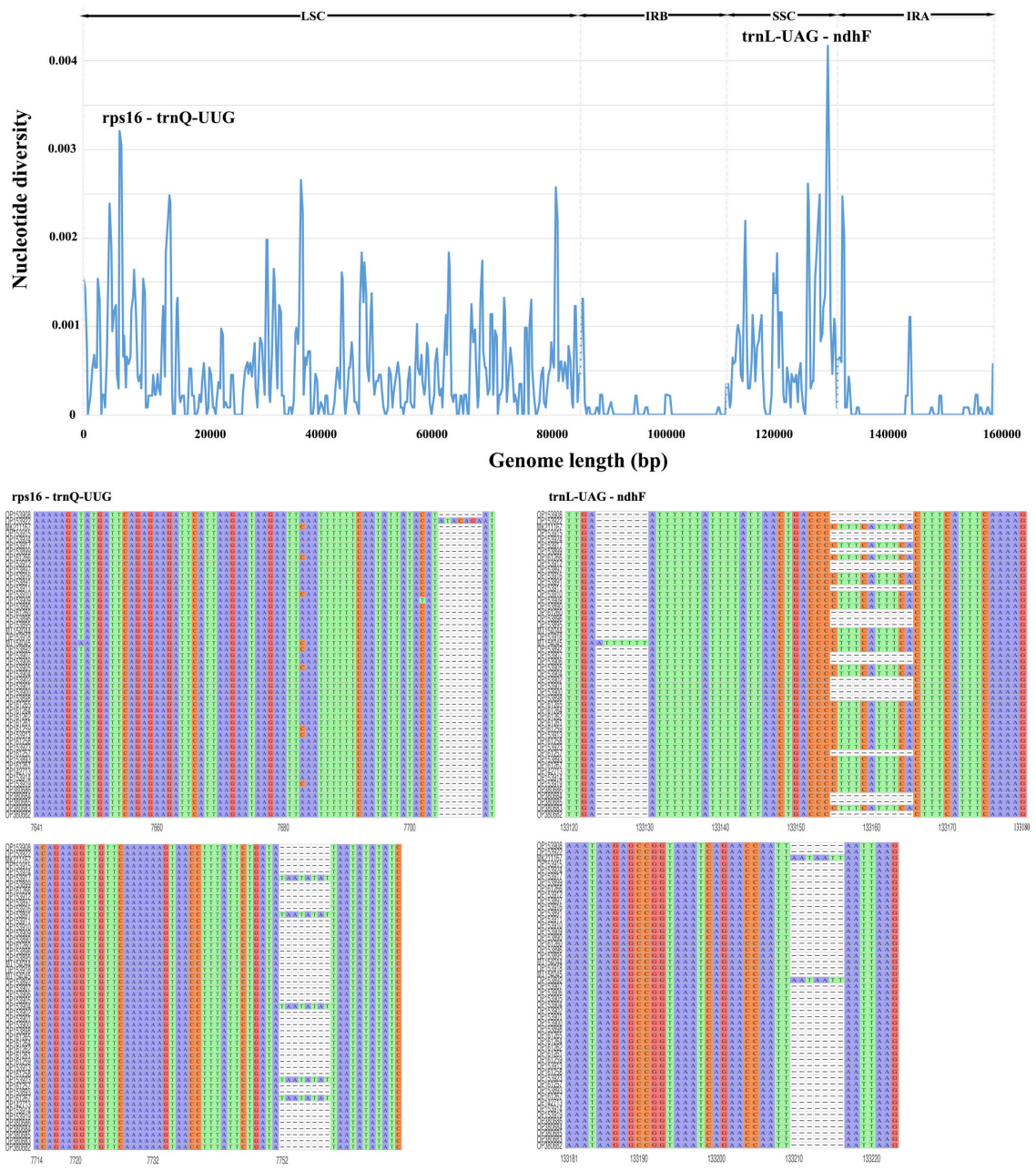


FIGURE 3
Comparative analysis of nucleotide variability by Pi values of the 50 CP genomes presented in a sliding window (window length: 500 bp; step size: 250 bp). X-axis: position of the midpoint of a window; Y-axis: nucleotide diversity in each window. The R package ggmsa was used to visualize the sequence alignment of two hotspot regions in the 50 CP genomes.

indicator of species affinity (Chen et al., 2021), and the GC content of *Morus* CP genomes showed slight differences, ranging from 36.13% to 36.21%, which indicated high conservation in the *Morus* CP genomes. Twenty-two intron-containing genes out of 125 genes were detected in these CP genomes. Among them, the trnK-UUU gene embodied the matK genes and had the largest intron (over 2,500 bp), which has been reported in previous studies (Li X. et al., 2019; Souza et al., 2020; Ren et al., 2022). matK is a well-known gene that is often used for molecular identification and analysis of genetic relationships in plants (Hilu and Liang, 1997; Ramesh et al., 2022), including *Morus* (Venkateswarlu et al., 2012). As a result, over one hundred matK genes of the *Morus* genus have been deposited in the GenBank of the NCBI. Intron-containing genes often have

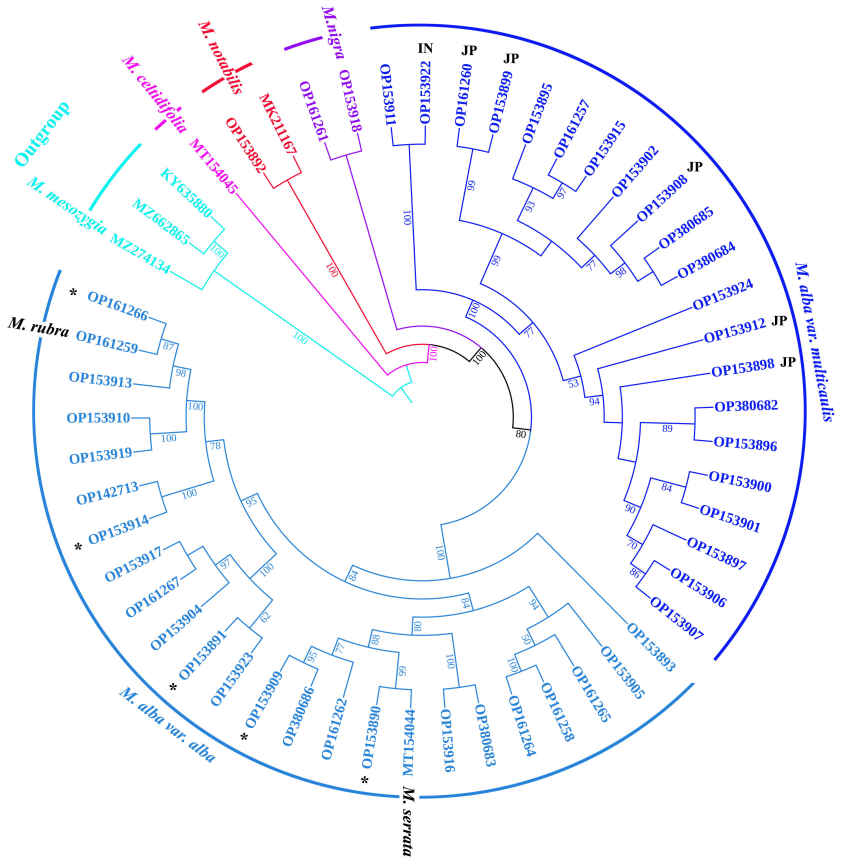


FIGURE 4
Phylogenetic relationships among *Morus* species based on their CP genomes with 50 *Morus* accessions and three outgroup genera (*Ficus*, *Broussonetia*, and *Afromorus*). The maximum-likelihood tree constructed with IQ-TREE2 is presented with complete CP genomes. The percentage of statistical support for the nodes is based on 1,000 bootstrap replicates. The black asterisks represent *Morus* accessions with mulberry fruit lengths over 4 cm. JP (Japan) and IN (India) in bold black represent mulberry accessions from Japan and India, respectively.

important physiological functions; for example, the clpP gene is relevant to proteolysis (Shikanai et al., 2001), and the ndhB gene has an important role in mediating photosystem I cyclic electron transport (Shen et al., 2022). Therefore, introns in *Morus* CP genomes may be useful in terms of physiological function.

The largest group with identical CP genome contained 28 mulberry germplasm accessions, including 9 Husang accessions and 9 Japanese mulberry accessions. Husang (or Hu mulberry, *M. alba* var. *multicaulis*, with multicaulis meaning many stalks or branches), a well-known cultivar of domesticated mulberry, is widely distributed worldwide (Kenrick, 1839; Jiao et al., 2020). Heyebai, named Husang-32, is a control cultivar in the National Mulberry New Cultivar Identification Test (Jiao et al., 2020). The selection of Husang germplasms was mainly performed for open-pollinated seedlings after the Song Dynasty, and the excellent traits were retained by the asexual method (Jiao et al., 2020). Tens of Husang germplasms were obtained and recorded after selection over hundreds of years. Most of the cultivated mulberry varieties in Japan are derived from the three original species, namely, Yamaguwa (*Morus bombycis*), Karayumaguwa (*M. alba*), and Roguwa (*Morus lhou*) (Minamizawa, 1997; Muhonja et al., 2020), of which *M. lhou* and *M. bombycis* (Koidzumi, 1930) belong to *M. alba* (Zeng et al., 2015). Additionally, Karayumaguwa and Roguwa were introduced to Japan from Chinese *M. alba* species around A.D. 677 and A.D. 1873, respectively (Hotta, 1958). Therefore, those three original species in Japan belong to the *M. alba* species. Ichibei, Kenmochi, and Gunmaakagi are related to Yamaguwa, whereas Ichinose, Kairyonezumigaeshi, Nezumigaeshi, and Kairyowasejumonji are related to Karayumaguwa (Minamizawa, 1997). In this study, 9 Japanese samples shared the same progenitor with Husang, indicating that the 9 Japanese samples belonged to *M. alba* var. *multicaulis*. Kairyonezumigaeshi was selected from among Nezumigaeshi plants in 1907 (Muhonja et al., 2020), which is consistent with their identical CP genomes (Table 1), whereas Ichinose (group EIGHT) was isolated from Nezumigaeshi (group ONE) seedlings in 1901 (Yamanouchi et al., 2010; Smethurst, 2014; Muhonja et al., 2020) and showed a different CP genome (Supplementary 1), which implied that hybridization may

increase the genetic diversity of CP genomes (Van Droogenbroeck et al., 2006; Daniell et al., 2016). In addition, in group EIGHT, 4 Japanese samples, Yamasou-159, Ichinose, Fengwei-Ichinose and Shinichinose, had common ancestors. Shinichinose, a hybrid variety derived from Ichinose (Muhonja et al., 2020), showed an identical CP genome with Ichinose, which is a typical character of maternal inheritance. In addition, group SIX showed another case of maternal inheritance because the maternal parent exhibited the same CP genome as its three hybrid progenies (Table 1 and Supplementary 1). The autotriploid cultivar Shaansang305 (group NINE, 159,200 bp) induced from the diploid cultivar Shinichinose using colchicine (Liu et al., 2021) showed obvious differences from the CP genome of Shinichinose (group EIGHT, 159,219 bp) (Supplementary 1), which suggested that polyploidizations affected the DNA of the nucleus and chloroplast (Choopeng et al., 2019; Zhai et al., 2021).

Group TWO consisted of 13 different accessions, including 12 *M. alba* var. *indica*, which indicated that they had common ancestors. Group THREE consisted of 9 different *M. notabilis* trees located in regions with an approximately 10 km radius of the pristine forest in Ya'an, Sichuan Province, Southwest China, and one seedling germinated from an *M. notabilis* seed. Group ELEVEN contained two different *M. yunnanensis* trees collected on Dawei Mountain, Yunnan Province, Southwest China.

The genome size (158,969-159,548 bp) and GC content (36.13%-36.21%) of the *Morus* CP genomes exhibited slight differences, which indicated high conservation of *Morus* CP genomes. DnaSP (Librado and Rozas, 2009) and mVISTA software were employed to investigate the divergence of CP genomes of the *Morus* genus. The results showed high conservation of gene order and rather low Pi values (0-0.00442), and noncoding regions were more variable than coding regions. At the same time, over half of the supermatrices showed identical sequences (Supplementary 2), which further supported that the coding regions were highly conserved. Therefore, the complete CP genomes of *Morus* were considered to construct the phylogenetic tree for identifying *Morus* species.

Phylogenetic analysis and taxonomical review of *Morus*

The 50 complete CP genomes of *Morus* were used to construct the ML tree for exploring the phylogenetic positions of the *Morus* species. It is clear that the genus *Morus* is monophyletic and is divided into five clades (Figure 4). Recently, *M. mesozygia* and *M. insignis*, native to Africa, which used to belong to the *Morus* genus, were eliminated from the *Morus* genus on the basis of phylogenetic analyses of supercontig sequences from 246 Moraceae samples (Gardner et al., 2021). Here, we also found that *M. mesozygia* did not belong to *Morus* because *M. mesozygia* was clustered with *Ficus* and *Broussonetia* (Figure 4).

Morus yunnanensis, similar to *M. notabilis* found in Southwest China with the same chromosome number, is a wild mulberry native to Yunnan Province, China (Xia et al., 2022). *Morus yunnanensis* was clustered with *M. notabilis* into a clade, indicating a close phylogenetic relationship (Figure 4). This close phylogenetic relationship between *M. yunnanensis* and *M. notabilis* was strongly supported by a phylogenomic tree (Xia et al., 2022). Combined with evidence from the nuclear genome and CP genome, we classify *M. yunnanensis* as belonging to *M. notabilis*.

Black mulberry (*M. nigra*), native to western Asia, is a Morus species with 308 chromosomes, which hinders the exploration of phylogenetic relationships based on the nuclear genome. It has been reported that *M. nigra* originated from *M. alba* (Agaev and Fedorova, 1970; Browicz, 2000; Lim., 2012), but molecular evidence is lacking. Fortunately, the CP genome is independent of the nuclear genome and is commonly used in phylogenetic studies. In our phylogenetic tree inferred from complete CP genomes, *M. nigra* displayed a close phylogenetic relationship with *M. alba* (Figure 4), which indicates that *M. nigra* originated from *M. alba*.

The largest clade in the phylogenetic tree of the *Morus* genus is the *M. alba* clade, comprising two subclades, *M. alba* var. *alba* and *M. alba* var. *multicaulis*, which is consistent with the taxonomy presented in the Flora of China (Zhou and Gilbert, 2003) and the phylogenetic tree based on domesticated mulberry accessions (Jiao et al., 2020). The *M. alba* var. *multicaulis* clade was divided into three subclades, including all Husang (*M. alba* var. *multicaulis*) and 16 Japanese mulberry samples, *M. alba* var. *indica* and other samples. Nine Japanese samples sharing the same CP genomes with Husang were clustered with seven other Japanese samples into the *M. alba* var. *multicaulis* clade, showing a close phylogenetic relationship, which indicated that these Japanese samples may have been derived from *M. alba* var. *multicaulis* through maternal inheritance. Recently, gene flow between Husang and Japanese samples was observed in the population structure analysis of mulberry accessions (Xia et al., 2022). Sixteen Japanese samples were clustered into two subclades, which was consistent with the findings of a previous report (Muhonja et al., 2020). Here, we provided molecular evidence that Japanese cultivated mulberry was derived from Chinese *M. alba* (Hotta, 1958; Jiao et al., 2020); therefore, we conclude that Japanese cultivated mulberry belongs to *M. alba* var. *multicaulis*. The subclade of *M. alba* var. *alba* contained red mulberry (*M. rubra*), *M. serrata*, *M. alba* var. *atropurpurea*, long-fruited mulberry germplasms, and other samples. Among them, *M. rubra*, a *Morus* species native to America, was clustered into *M. alba* var. *alba*, which may be triggered by common hybridization with *M. alba* (Burgess et al., 2008). It has been

reported that *M. rubra* commonly hybridizes with *M. alba* and that *M. alba* potentially poses a threat to the existence of *M. rubra*, which leads to the endangerment of native *M. rubra* in America (Nepal and Wichern, 2013). In field observations, the direction of introgression of hybrids between *M. rubra* and *M. alba* was biased toward *M. alba* as the maternal parent (Nepal and Wichern, 2013). Mulberry germplasms with long fruits include *M. wittiorum* and *M. macroura*, which have been recognized as *M. alba* (Zeng et al., 2015). Here, we supplied new molecular evidence at the genome level. *Morus serrata* was classified as a species based on morphological taxonomy and molecular marker genes (Hotta, 1958; Zhou and Gilbert, 2003; Nepal and Ferguson, 2012; Zeng et al., 2015). However, in this study, *M. serrata* was clustered with *M. alba* var. *alba* in the phylogenetic tree based on the CP genome (Figure 4). The classification of *M. serrata* requires more samples and further investigation using molecular evidence.

Indian mulberry (*M. indica* or *M. alba* var. *indica*) is recognized as a variety of *M. alba* (Datta, 1954; Rao and Jarvis, 1986; Zeng et al., 2015; Muhonja et al., 2020). The first *Morus* CP genome (GenBank: DQ226511, 158,484 bp) was identified in *M. indica* with the reference genome method using plastid genomic DNA (Ravi et al., 2006). In this study, the CP genomes of 12 different samples, including *M. indica* var. kanva2, were *de novo* assembled and found to have identical sequences (size: 158,969 bp). Sample SL2 in group TWO, native to Sri Lanka (Xia et al., 2022), may be a hybrid progeny of *M. alba* var. *indica*. *M. indica* was clustered with *M. alba* var. *multicaulis*, indicating a close phylogenetic relationship with this taxon (Figure 4). Additional molecular evidence of the phylogenetic relationships of *M. indica* was provided (Muhonja et al., 2020). Therefore, *M. indica* may be derived from *M. alba* var. *multicaulis*.

In the present study, we *de novo* assembled 123 *Morus* CP genomes and found that they are highly conserved. Many *Morus* CP genomes displayed identical sequences, which indicated that they shared common maternal ancestors. We propose that the *Morus* genus includes six species, namely, *M. notabilis*, *M. celtidifolia*, *M. nigra*, *M. rubra*, *M. serrata*, and *M. alba* comprising two subspecies, *M. alba* var. *alba* and *M. alba* var. *multicaulis*. The Japanese cultivated germplasms were derived from *M. alba* var. *multicaulis*. Our findings provide valuable information for studies on the classification, domestication, and breeding improvement of mulberry.

Data availability statement

The raw sequence data were deposited in the CNGB Sequence Archive of the China National GeneBank Database (CNGBdb) under accession number CNP 0001407.

Author contributions

QZ, NH, and ZX conceived the project and designed the experiments. QZ assembled, annotated, and analyzed the genomes, QZ wrote the manuscript. MC, SW and XX contributed materials and isolated DNA and analyzed the data. TL helped QZ to analyze the data. All authors contributed to the article and approved the submitted version.

Funding

This work was funded by the National Key Research and Development Program (No. 2018YFD1000602) and the Chongqing Research Program of Basic Research and Frontier Technology (cstc2021yszx-jcyj0004).

Acknowledgments

We are grateful to Professor Elizabeth Makings from Arizona State University for kindly providing leaves of *M. celtidifolia*. We thank for the help of Ni Yang and Li Jingling from the Institute of Medicinal Plant Development for their suggestion on the annotation and analysis of chloroplast genomes. We also thank the reviewers for helpful comments on 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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1047592/full#supplementary-material

References

Agaev, Y. M., and Fedorova, H. E. (1970). Investigation of meiosis in the diploid species moras alba l., the 22-ploid m. nigra l. and their cross in relation to the origin of the species m. nigra l. *Pak. J. Bot.* 2 (1), 65–76.

Browicz, K. (2000). *Where is the place of origin of morus nigra (Moraceae)? Fragm. Flor. Geobot.* 45 (1/2), 273–280.

Bruun-Lund, S., Clement, W. L., Kjellberg, F., and Rønsted, N.. (2017). First plastid phylogenomic study reveals potential cyto-nuclear discordance in the evolutionary history of ficus L.(Moraceae). *Mol. Phylogenet. Evol.* 109, 93–104. doi: 10.1016/j.ympev.2016.12.031

Bureau, L.É. (1873). "Moraceae," in *Prodromus systematis naturalis regni vegetabilis*. Ed. A. P. DeCandolle (Paris, France: Tuettel and Wurtz), 211–288.

Burgess, K. S., Martin, M., and Husband, B. C. (2008). *Interspecific seed discounting and the fertility cost of hybridization in an endangered species. New Phytol.* 177 (1), 276–283. doi: 10.1111/j.1469-8137.2007.02244.x

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinf. (Oxford England)* 25 (15), 1972–1973. doi: 10.1093/bioinformatics/btp348

Chen, C., Zhou, W., Huang, Y., and Wang, Z. Z. (2016). The complete chloroplast genome sequence of the mulberry morus notabilis (Moreae). *Mitochondrial DNA A DNA Mapp. Seq. Anal.* 27 (4), 2856–2857. doi: 10.3109/19401736.2015.1053127

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

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

Choopeng, S., Te-Chato, S., and Khawnium, T. (2019). Effect of colchicine on survival rate and ploidy level of hydrid between dendrobium santana x d. friedericksianum orchid. *Int. J. Agric. Technol.* 15 (2), 249–260.

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

Datta, M. (1954). Cytogenetical studies on two species of morus. *Cytologia* 19 (1), 86–95. doi: 10.1508/cytologia.19.86

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

Gardner, E. M., Garner, M., Cowan, R, Dodsworth, S, Epitawalage, N, Arifiani, D, et al. (2021). Repeated parallel losses of inflexed stamens in moraceae: Phylogenomics and generic revision of the tribe moreae and the reinstatement of the tribe olmedieae (Moraceae). *Taxon.* 70, 946 988. doi: 10.1002/tax.12526

Gitzendanner, M. A., Soltis, P. S., Wong, G. K. S., Ruhfel, B. R., Soltis, D. E., et al. (2018). Plastid phylogenomic analysis of green plants: a billion years of evolutionary history. *Am. J. Bot.* 105 (3), 291–301. doi: 10.1002/ajb2.1048

Group A. P. (2009). An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG III. *Bot. J. Linn. Soc.* 161 (2), 105–121. doi: 10.1111/j.1095-8339.2009.00996.x

He, N., Zhang, C., Qi, X., Zhao, S., Tao, Y., Yang, G., et al. (2013). Draft genome sequence of the mulberry tree morus notabilis. *Nat. Commun.* 4, 2445. doi: 10.1038/ncomms3445

He, S.-L., Tian, Y., Yang, Y., and Shi, C. -Y. (2020). *Chloroplast genome and phylogenetic analyses of morus alba (Moraceae). Mitochondrial DNA Part B* 5 (3), 2203–2204. doi: 10.1080/23802359.2019.1673242

Hilu, K. W., and Liang, G. (1997). The matK gene: sequence variation and application in plant systematics. *Am. J. Bot.* 84 (6), 830–839. doi: 10.2307/2445819

Hotta, T. (1954). Fundamentals of morus plants classification. *Kinugasa Sanpo* 390, 13–21.

Hotta, T. (1958). *Taxonomical studies on the morus plants and their distributions in Japan and its vicinities* (Tokyo: Japan Society for the Promotion of Science).

Hua, Z., Tian, D., Jiang, C., Song, S., Chen, Z., Zhao, Y., et al. (2022). *Towards comprehensive integration and curation of chloroplast genomes. Plant Biotechnol. J.* doi: 10.1111/pbi.13923

Jain, M., Bansal, J., Rajkumar, M. S., Sharma, N., Khurana, J. P., and Khurana, P. (2022). Draft genome sequence of Indian mulberry (Morus indica) provides a resource for functional and translational genomics. *Genomics* 114 (3), 110346. doi: 10.1016/j.ygeno.2022.110346

Jiao, F., Luo, R., Dai, X., Lui, H., Yu, G., Han, S., et al. (2020). Chromosome-level reference genome and population genomic analysis provide insights into the evolution and improvement of domesticated mulberry (Morus alba). *Mol. Plant* 13 (7), 1001–1012. doi: 10.1016/j.molp.2020.05.005

Jin, J.-J., Yu, W. -B., Yang, J. B., Song, Y., Depamphilis, C. W., Yi, Y. -S., et al. (2020). GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* 21 (1), 1–31. doi: 10.1186/s13059-020-02154-5

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

Kenrick, W. (1839). *The American silk grower's guide : or, the art of raising the mulberry and silk, and the system of successive crops in each season. 2d ed* (Boston: Weeks, Jordan & co).

Koidzumi, G. (1930). *Florae symbolae orientali-asiaticae: Sive, contributions to the knowledge of the flora of Eastern Asia* (Kyoto, Japan) 115 pp. doi: 10.11501/167884

Kong, W., and Yang, J. (2016). The complete chloroplast genome sequence of morus mongolica and a comparative analysis within the fabidae clade. *Curr. Genet.* 62 (1), 165–172. doi: 10.1007/s00294-015-0507-9

Kong, W. Q., and Yang, J. H. (2017). The complete chloroplast genome sequence of morus cathayana and morus multicaulis, and comparative analysis within genus morus l. *PeerJ* 5, e3037. doi: 10.7717/peerj.3037

Letunic, I., and Bork, P. (2019). *Interactive tree of life (iTOL) v4: recent updates and new developments. Nucleic Acids Res.* 47 (W1), W256–W259. doi: 10.1093/nar/gkz239

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

Li, Q. L., Guo, J. Z., Yan, N., and Li, C. C. (2016). Complete chloroplast genome sequence of cultivated morus l. species. *Genet. Mol. Res.* 15 (4), 1–13. doi: 10.4238/gmr15048906

Li, H.-T., Yi, T. -S., Gao, L. -M., Ma, P. -F., Zhang, T., Yang, J. B., et al. (2019). Origin of angiosperms and the puzzle of the Jurassic gap. *Nat. Plants* 5 (5), 461–470. doi: 10.1038/s41477-019-0421-0

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), 1045. doi: 10.3390/ijms20051045

Li, H.-T., Luo, Y., Gan, L., Ma, P. -F., Gao, L. -M., Yang, J. -B., et al. (2021). Plastid phylogenomic insights into relationships of all flowering plant families. *BMC Biol.* 19 (1), 1–13. doi: 10.1186/s12915-021-01166-2

Lim., T. K. (2012). "Morus nigra, in edible medicinal and non-medicinal plants," in *Fruits*, vol. Volume 3. (Netherlands: Springer), 430–438.

Linnaeus, C. "Morus," in *Species plantarum* (1753). (Stockholm, Impensis Laurentii Salvii) 968. doi: 10.5962/bhl.title.669

Liu, H., Sun, H., Bao, L., Han, S., Hui, T., Zhang, R., et al. (2021). Secondary metabolism and hormone response reveal the molecular mechanism of triploid mulberry (Morus alba l.) trees against drought. *Front. Plant Sci.* 12, 720452. doi: 10.3389/fpls.2021.720452

Lohse, M., Drechsel, O., Kahlau, S., and Bock, P.. (2013). *OrganellarGenomeDRAW–a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Res.* 41 (Web Server issue), W575. doi: 10.1093/nar/gkt289

Lowe, T. M., and Chan, P. P. (2016). *tRNAscan-SE on-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res.* 44, W54–W57. doi: 10.1093/nar/gkw413

Luo, J., Wang, Y., and Zhao, A. Z. (2019). *The complete chloroplast genome of morus alba (Moraceae: Morus), the herbal medicine species in china. Mitochondrial DNA Part B* 4 (2), 2467–2468. doi: 10.1080/23802359.2019.1638328

Minamizawa, K. (1997). *Moriculture: Science of mulberry cultivation. 2 ed* (Boca Raton, Florida, USA: CRC Press).

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., et al. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37 (5), 1530–34. doi: 10.1093/molbev/msaa015

Moore, M. J., Bell, C. D., Soltis, P. S., and Soltis, D. E. (2007). Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proc. Natl. Acad. Sci.* 104 (49), 19363–19368. doi: 10.1073/pnas.0708072104

Moore, M. J., Soltis, P. S, Bell, C. D., Burleigh, J. D., Soltis, D. E., et al. (2010). Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proc. Natl.Acad. Sci. U.S.A* 107 (10), 4623–4628. doi: 10.1073/pnas.0907801107

Muhonja, L., Yamanouchi, H., Yang, C. -C., Kuwazaki, S., Yokoi, K., Kameda, T., et al. (2020). Genome-wide SNP marker discovery and phylogenetic analysis of mulberry varieties using double-digest restriction site-associated DNA sequencing. *Gene.* 726, 144162. doi: 10.1016/j.gene.2019.144162

Nepal, M. P., and Ferguson, C. J. (2012). Phylogenetics of morus (Moraceae) inferred from ITS and trnL-trnF sequence data. *Syst. Bot.* 37 (2), 442–450. doi: 10.1600/036364412X635485

Nepal, M. P., and Wichern, D. J. (2013). Taxonomic status of red mulberry (morus rubra, moraceae) At its northwestern boundar. *Proc. South Dakota Acad. Sci.* 92, 19–29.

Ramesh, G. A., Mathew, D., John, K. J., and Ravisankar, V. (2022). Chloroplast gene matK holds the barcodes for identification of momordica (Cucurbitaceae) species from Indian subcontinent. *Hortic. Plant J.* 8 (1), 89–98. doi: 10.1016/j.hpj.2021.04.001

Rao, C. K., and Jarvis, C. (1986). Lectotypification, taxonomy and nomenclature of morus alba, m. tatarica and m. indica (Moraceae). *Taxon.* 35 (4), 705–708. doi: 10.2307/1221619

Ravi, V., Khurana, J. P., Tyagi, A. K., and Khurana, P. (2006). The chloroplast genome of mulberry: complete nucleotide sequence, gene organization and comparative analysis. *Tree Genet. Genomes* 3 (1), 49–59. doi: 10.1007/s11295-006-0051-3

Ren, J., Tian, J., Jiang, H., Zhu, X. -X., Mutie, F. M., Wang, F. M., 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, 828467. doi: 10.3389/fpls.2022.828467

SB, D., Dhar, A., and Sengupta, K. (1990). Meiosis in natural decosaploid (22x) morus nigra l. *Cytologia* 55 (3), 505–509.

SB, D., and Rajan, M. V. (1989). Microsporogenesis in hexaploid morus serrata roxb. *Cytologia* 54 (4), 747–751.

Scheunert, A., Dorfner, M., Lingl, T., and Oberprieler, C. (2020). Can we use it? on the utility of de novo and reference-based assembly of nanopore data for plant plastome sequencing. *PloS One* 15 (3), e0226234. doi: 10.1371/journal.pone.0226234

Shen, L., Tang, K., Wang, W., Wang, C., Wu, H., Mao, Z., et al. (2022). Architecture of the chloroplast PSI–NDH supercomplex in hordeum vulgare. *Nature* 601 (7894), 649–654. doi: 10.1038/s41586-021-04277-6

Shi, L., Chen, H., Jiang, M., Wang, M, 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

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 (3), 264–273. doi: 10.1093/pcp/pce031

Smethurst, R. J. (2014). *Agricultural development and tenancy disputes in Japan* (Princeton, NJ: Princeton University Press) 1870–1940.

Souza, U. J. B. D., Vitorino, L. C., Bessa, L. A., Silva, F. G., et al. (2020). The complete plastid genome of artocarpus camansi: a high degree of conservation of the plastome structure in the family moraceae. *Forests* 11 (11), 1179. doi: 10.3390/f11111179

Stull, G. W., Duno De Stefano, R., Soltis, D. E., and Soltis, P. S. (2015). Resolving basal lamiid phylogeny and the circumscription of icacinaceae with a plastome-scale data set. *Am. J. Bot.* 102 (11), 1794–1813. doi: 10.3732/ajb.1500298

Sun, Y., Moore, M. J., Zhang, S., Soltis, P. S., Soltis, D. E., Zhao, T., et al. (2016). Phylogenomic and structural analyses of 18 complete plastomes across nearly all families of early-diverging eudicots, including an angiosperm-wide analysis of IR gene content evolution. *Mol. Phylogenet. Evol.* 96, 93–101. doi: 10.1016/j.ympev.2015.12.006

Tikader, A., and Kamble, C. K. (2008). Mulberry wild species in India and their use in crop improvement–a review. *Aust. J. Crop Sci.* 2 (2), 64–72.

Van Droogenbroeck, B., Kyndt, T., Romeijn-Peeters, E., Van Thuyne, W, Goetghebeur, P., Romero-Motochi, J., et al. (2006). Evidence of natural hybridization and introgression between vasconcellea species (Caricaceae) from southern Ecuador revealed by chloroplast, mitochondrial and nuclear DNA markers. *Ann. Bot.* 97 (5), 793–805. doi: 10.1093/aob/mcl038

Venkateswarlu, M., Ravikumar, G., Vijayaprakash, N., Rao, C., Kamble, C., and Tikadar, A. (2012). Molecular phylogeny of morus species differentiation based on chloroplast matK sequences. *Indian J. Sericulture* 51 (1), 16–19.

Vijayan, K. S., and B.da Silva, J. A. T. (2011). Germplasm conservation in mulberry (Morus spp.). *Scientia Hortic.* 128 (4), 371–379. doi: 10.1016/j.scienta.2010.11.012

Vijayan, K., Tikader, A., Weiguo, Z., Nair, C. V., Ercisli, S., and Tsou, C. -H. (2011). "Morus," in *Wild crop relatives: Genomic and breeding resources (Tropical and subtropical fruits)*. Ed. K. Chittaranjan (Berlin Heidelberg: Springer-Verlag), 75–95.

Wang, G., Zhang, X., Herre, E. A., Mckey, D., Machado, C. A., Yu, W. -B., et al. (2021). Genomic evidence of prevalent hybridization throughout the evolutionary history of the fig-wasp pollination mutualism. *Nat. Commun.* 12 (1), 1–14. doi: 10.1038/s41467-021-20957-3

Wang, W., and Lanfear, R. (2019). Long-reads reveal that the chloroplast genome exists in two distinct versions in most plants. *Genome Biol. Evol.* 11 (12), 3372–3381. doi: 10.1093/gbe/evz256

Xia, Z., Dai, X., Fan, W., Liu, Z., Zhang, M., Bian, P., et al. (2022). Chromosome-level genomes reveal the genetic basis of descending dysploidy and sex determination in morus plants. *Genomics Proteomics Bioinf.* doi: 10.1016/j.gpb.2022.08.005

Xuan, Y., Wu, Y., Li, P., Liu, R., Luo, Y., Yuan, J., et al. (2019). Molecular phylogeny of mulberries reconstructed from ITS and two cpDNA sequences. *PeerJ.* 7, e8158. doi: 10.7717/peerj.8158

Xuan, Y., Ma, B., Li, D., Tian, Y., Zeng, Q., and He, N. (2022). Chromosome restructuring and number change during the evolution of *Morus notabilis and Morus alba*. *Hortic. Res.* 9, uhab030. doi: 10.1093/hr/uhab030

Yamanouchi, H., Koyama, A., Takyu, T., and Muramatsu, N. (2010). Nuclear DNA amounts in diploid mulberry speceis (Morus. *spp.) J. Insect Biotechnol. Sericology* 79, 1–8.

Zeng, Q., Chen, H., Zhang, C., Han, M., Li, T., Qi, X., et al. (2015). Definition of eight mulberry species in the genus morus by internal transcribed spacer-based phylogeny. *PloS One* 10 (8), e0135411. doi: 10.1371/journal.pone.0135411

Zhai, Y., Yu, X., Zhou, J., Li, J., Tian, Z., and Wang, P. (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, Z.-R., Yang, X., Li, W. -Y., Peng, Y. -Q., and Gao, J. (2022). Comparative chloroplast genome analysis of ficus (Moraceae): Insight into adaptive evolution and mutational hotspot regions. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.965335

Zhou, L., Feng, T., Xu, S., Gao, F., Lam, T. T., Wang, Q., et al. (2022). Ggmsa: a visual exploration tool for multiple sequence alignment and associated data. *Briefings Bioinf.* doi: 10.1093/bib/bbac222

Zhou, Z., and Gilbert, M. G. (2003). "Moraceae," in *Flora of China*. Eds. Z. Y. Wu, P. H. Raven and D. Y. Hong (Beijing, China & Saint Louis, Missouri: Science Press & Missouri Botanical Garden Press), 22–26.

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 01 December 2022
DOI 10.3389/fpls.2022.1045423

OPEN ACCESS

EDITED BY
Lianming Gao,
Kunming Institute of Botany, Chinese Academy of Sciences (CAS), China

REVIEWED BY
Lihong Xiao,
Zhejiang Agriculture and Forestry University, China
Hanghui Kong,
South China Botanical Garden, Chinese Academy of Sciences (CAS), China

*CORRESPONDENCE
Linchun Shi
linchun_shi@163.com
Xiaoxia Zhang
zhangxiaoxia@ibcas.ac.cn

†These authors have contributed equally to this work

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

RECEIVED 15 September 2022
ACCEPTED 10 November 2022
PUBLISHED 01 December 2022

CITATION
Liu J, Shi M, Zhang Z, Xie H, Kong W, Wang Q, Zhao X, Zhao C, Lin Y, Zhang X and Shi L (2022) Phylogenomic analyses based on the plastid genome and concatenated nrDNA sequence data reveal cytonuclear discordance in genus *Atractylodes* (Asteraceae: Carduoideae).
Front. Plant Sci. 13:1045423.
doi: 10.3389/fpls.2022.1045423

Phylogenomic analyses based on the plastid genome and concatenated nrDNA sequence data reveal cytonuclear discordance in genus *Atractylodes* (Asteraceae: Carduoideae)

Jinxin Liu[1†], Mengmeng Shi[1,2†], Zhaolei Zhang[1,2], Hongbo Xie[2], Weijun Kong[3], Qiuling Wang[1], Xinlei Zhao[1], Chunying Zhao[2], Yulin Lin[1], Xiaoxia Zhang[4*] and Linchun Shi[1*]

[1]Key Laboratory of Chinese Medicine Resources Conservation, State Administration of Traditional Chinese Medicine of the People's Republic of China, Engineering Research Center of Chinese Medicine Resource of Ministry of Education, Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China, [2]Hebei Key Laboratory of Study and Exploitation of Chinese Medicine, Chengde Medical University, Chengde, China, [3]School of Traditional Chinese Medicine, Capital Medical University, Beijing, China, [4]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing, China

Atractylodes species are widely distributed across East Asia and are cultivated as medicinal herbs in China, Japan, and Korea. Their unclear morphological characteristics and low levels of genetic divergence obscure the taxonomic relationships among these species. In this study, 24 plant samples were collected representing five species of *Atractylodes* located in China; of these, 23 belonged to members of the *A. lancea* complex. High-throughput sequencing was used to obtain the concatenated nrDNA sequences (18S-ITS1-5.8S-ITS2-28S) and plastid genomes. The concatenated nrDNA sequence lengths for all the *Atractylodes* species were 5,849 bp, and the GC content was 55%. The lengths of the whole plastid genome sequences ranged from 152,138 bp (*A. chinensis*) to 153,268 bp (*A. lancea*), while their insertion/deletion sites were mainly distributed in the intergenic regions. Furthermore, 33, 34, 36, 31, and 32 tandem repeat sequences, as well as 30, 30, 29, 30, and 30 SSR loci, were detected in *A. chinensis*, *A. koreana*, *A. lancea*, *A. japonica*, and *A. macrocephala*, respectively. In addition to these findings, a considerable number of heteroplasmic variations were detected in the plastid genomes, implying a complicated phylogenetic history for *Atractylodes*. The results of the

Abbreviations: IR, inverted repeat region; LSC, large single copy region; SSC, small single copy region; tRNA, transfer RNA; rRNA, ribosomal RNA; SSR, simple sequence repeats; ML, Maximum Likelihood.

phylogenetic analysis involving concatenated nrDNA sequences showed that *A. lancea* and *A. japonica* formed two separate clades, with *A. chinensis* and *A. koreana* constituting their sister clade, while *A. lancea*, *A. koreana, A. chinensis*, and *A. japonica* were found based on plastid datasets to represent a mixed clade on the phylogenetic tree. Phylogenetic network analysis suggested that *A. lancea* may have hybridized with the common ancestor of *A. chinensis* and *A. japonica*, while ABBA–BABA tests of SNPs in the plastid genomes showed that *A. chinensis* was more closely related to *A. japonica* than to *A. lancea*. This study reveals the extensive discordance and complexity of the relationships across the members of the *A. lancea* complex (*A. lancea, A. chinensis*, *A. koreana*, and *A. japonica*) according to cytonuclear genomic data; this may be caused by interspecific hybridization or gene introgression.

KEYWORDS

***Atractylodes*, plastid genome, nrDNA concatenated sequence, phylogeny, cyto-nuclear discordance**

Introduction

Atractylodes, mainly distributed across East Asia, is a perennial herb from the Asteraceae family. The dried rhizomes of *Atractylodes* plants have been used for more than 2,000 years as traditional herbal medicines ("cangzhu" and "baizhu" in China, and so-jutsu and byaku-jutsu in Japan) to treat gastroduodenal diseases and colds (Yin et al., 2015; Deng et al., 2016). Due to the limitations of wild resources, the plants of *Atractylodes* have been cultivated in China since the 1980s, and production of their dried rhizome has now reached 5000 tons per year. However, the cultivation of particular species of *Atractylodes* presents the challenge of heterogeneous germplasm, which is mainly caused by high variability and continuous variation in morphological features among individual plants (Deng et al., 2016). In 1959, *A. chinensis* DC. was considered an independent species in Northeast Medicinal Flora (Liu, 1959) and was divided into several variants, such as *A. chinensis* var. *koreana* (Nakai) Chu, *A. chinensis* DC. var. *simplicifolia* (Loes.) Chu, and *A. chinensis* DC. var. *liaotungensis* Kitag. In 1981, Shi (Zhu, 1981) considered pinnatipartite leaves to be a volatile mutation, and further claimed, following an in-depth analysis of documented specimens and the literature, that the labels *A. chinensis* and *A. lancea* were actually used to refer to a single species. In 1987, *A. chinensis* was treated as a synonym of *A. lancea* according to the law of priority nomenclature in Flora Reipublicae Popularis Sinicae (Lin and Shi, 1987). In 2011, according to Flora of China, *A. japonica* was also treated as a synonym of *A. lancea;* at present, *A. lancea* is an overcrowded group whose leaves are described as undivided or divided almost to base into 3-5(-9) pinnately arranged segments (Shi, 2011). The nomenclatural history of *Atractylodes* is illustrated in Supplementary Figure 1. Although several molecular markers have been used to investigate the taxonomic and phylogenetic characteristics of *Atractylodes* (Peng et al., 2012; Kim et al., 2016), there is still taxonomic controversy surrounding this genus, arising primarily from the interspecific and intraspecific taxonomic treatment of the *A. lancea* complex, consisting of *A. japonica*, *A. koreana*, *A. lancea*, and *A. chinensis*.

The rapid development of next-generation sequencing (NGS) technology (McPherson et al., 2013), coupled with powerful bioinformatic tools (Liu et al., 2012; Shi et al., 2019), has made it possible to study the genomic evolution and interspecific relationships of organisms according to whole plastid genomic data. Substantial nucleotide substitution, indel events, and structural rearrangements have been found in plastid genomes, indicating that the whole plastid genome contains a significant amount of phylogenetic information (Liu et al., 2018). Additionally, plant plastids display intraspecific heteroplasmic variation, a term which refers to the presence of nonidentical plastid molecules in a cell or organism (Scarcelli et al., 2016). Previous studies have used NGS methods to detect heteroplasmy in the plastid genome of *Astragalus membranaceus*, which could explain why the *de novo* genome assembly program has failed to assemble the genome in heterogeneous regions (Lei et al., 2016). Moreover, intra-individual polymorphism can provide new evidence that can be used in evolutionary and classification analysis (Sun et al., 2019). In addition, extensive phylogenetic discordance among nuclear and organellar phylogenies has been found in the genus *Sphagnum*; this has been caused by incomplete lineage sorting (ILS) following the rapid radiation of the genus, rather than by post-speciation introgression (Meleshko et al., 2021).

In 2020, Wang et al. used the plastomes and nuclear sequences of *Atractylodes lancea*, *A. chinensis*, and *A. macrocephala* to reconstruct the phylogenetic relationships of these three species (Wang et al., 2020). Phylogeny analysis using the plastid data indicated that *A. lancea* and *A. chinensis* are more closely related to one another than to *A. macrocephala*. Interestingly, this study further observed intra-individual polymorphism of SLD5, such that SLD5 has two haplotypes in *A. macrocephala*, of which one can be found in *A. lancea* and, separately, the other can be found in *A. chinensis*. This intra-individual polymorphism was taken to imply that *A. macrocephala* may be a hybrid of *A. lancea* and *A. chinensis*, or the result of introgressive hybridization (Wang et al., 2020). Subsequently, analysis of six plastid genomes of *Atractylodes* (*A. chinensis*, *A. koreana*, *A. lancea*, *A. macrocephala*, *A. japonica*, and *A. carlinoides*) has revealed that the phylogenetic relationship within *Atractylodes* is complex (Wang et al., 2021). The results indicated that *A. japonica* and *A. lancea* are clustered into a subclade, while *A. chinensis* and *A. koreana* are clustered into another subclade. The abovementioned studies have confirmed that plastid genome analysis is a valuable tool for the phylogenetic study of *Atractylodes*, and additional specimens should be collected to obtain further evidence on the complex evolutionary history of *A. lancea* (Wang et al., 2021).

In this study, we collected 24 plant samples, 23 of which represented species of *A. lancea* complex. High-throughput sequencing was used to obtain the concatenated nrDNA sequences (18S-ITS1-5.8S-ITS2-28S) and plastid genomes. Our analyses showed that *A. chinensis* was more closely related to *A. japonica* than to *A. lancea*. Furthermore, extensive discordance and complex relationships across the genus of *Atractylodes* were revealed through analysis of the cytonuclear genomic data.

Materials and methods

Sample collection

For this study, 24 samples representing specimens of *Atractylodes lancea*, *A. chinensis*, *A. macrocephala*, *A. japonica*, and *A. koreana* were collected, of which six samples were collected from wild regions and 18 samples from cultivated regions (Supplementary Figure 2). All the samples were morphologically authenticated by Chunying Zhao (Chengde Medical University), Xinlei Zhao (Institute of Medicinal Plant Development, CAMS), Yulin Lin (Institute of Medicinal Plant Development, CAMS), and Qiuling Wang (Institute of Medicinal Plant Development, CAMS). Detailed information is provided in Supplementary Table 1 and Supplementary Figure 3. In addition, data relating to 20 samples representing six species of genus *Atractylodes* and ten species of outgroup taxa were downloaded from GenBank (Supplementary Table 2).

DNA extraction, library preparation, and high-throughput sequencing

Total genomic DNA extraction was performed on the leaf tissues using a modified CTAB method. The quantity and quality of the DNA were determined using Qubit 4.0 (Thermo Fisher Scientific Inc., USA). The sequencing library (~350 bp) was constructed using purified DNA and a TruSeq DNA PCR-Free High Throughput Library Prep Kit (Illumina USA). An Illumina NovaSeq platform was employed to conduct high-throughput sequencing. The raw data were deposited in the Sequence Read Archive (SRA) under BioProject accession number PRJNA682118. The final plastid genomes and concatenated nrDNA sequences of the *A. lancea*, *A. chinensis*, *A. macrocephala*, *A. japonica*, and *A. koreana* specimens were assembled, annotated, and submitted to GenBank (Supplementary Table 1).

Assembly, annotation, and characterization of the concatenated nrDNA and plastid genome sequences

The sequencing adapter and low-quality reads were filtered using Trimmomatic v0.38 (Bolger et al., 2014). Whole plastid genomes were assembled *via* the organelle assembler NOVOPlasty v4.2.1 (Jin et al., 2020) and GetOrganelle (Jin et al., 2020). The plastid genome sequence of *A. lancea* (accession number: NC_037483) was selected as a reference in the NOVOPlasty configuration file. CpGAVAS2 (www.herbalgenomics.org/cpgavas2) with default parameters was used to annotate the protein-coding, rRNA, and tRNA genes of the plastid genome and to facilitate visualization (Shi et al., 2019), with the initial annotations being edited manually using the Apollo genome editor. A circular map was generated using OrganellarGenomeDRAW (OGDRAW) (Greiner et al., 2019). The concatenated nrDNA sequences (18S, ITS1, 5.8S, ITS2, and 28S nrDNA) were assembled using Getorganelle and compared with the nuclear ribosomal RNA database to obtain the annotation results. The codon usage and relative synonymous codon usage (RSCU) of the plastid genomes were calculated using CodonW (http://codonw.sourceforge.net/).

Analysis of repeat structures and intraspecific variation in the plastid genomes

The REPuter (Kurtz et al., 2001) program was used to identify four types of sequence repeats, including forward (F), reverse (R), complementary (C), and palindromic (P). The minimum repeat size for oligonucleotide repeats was set at 30

bp, with a Hamming distance of 3 (i.e., a sequence identity of 90%). Tandem repeats were analyzed using the TRF (Benson, 1999) software with default parameters. Simple sequence repeats (SSRs) were detected using the MIcroSAtellite identification tool (MISA, available online: http://pgrc.ipk-gatersleben.de/misa/) (Beier et al., 2017) with minimum repeat thresholds of 10, 6, 5, 5, 5, and 5 for mono-, di-, tri-, tetra-, penta-, and hexa-nucleotide SSRs, respectively. Intraspecific variations were detected by first mapping the reads to reference sequences using Bowtie2 (Langmead and Salzberg, 2012), and subsequently conducting analysis using our local python program and visualizing the sequences using an integrative genomics viewer.

Comparative analysis of the plastid genomes

The mVISTA program (http://genome.lbl.gov/vista/mvista/submit.shtml) was used in Shuffle-LAGAN mode for the comparative analysis of divergence regions with default parameters, and *A. chinensis* was used as a reference. Intra- and inter-distances were analyzed in accordance with procedures reported in our previous study (Chen et al., 2010). Mauve (Darling et al., 2004), a system used to construct multiple genome alignments in the presence of large-scale evolutionary events, was used to identify locally collinear blocks (LCBs) of *Atractylodes* species. The contraction and expansion of the IR boundaries between the four main parts of the genome (LSC/IRb/SSC/IRa) were visualized using IRscope (https://irscope.shinyapps.io/irapp/).

Phylogenetic and gene introgression analysis

A total of 44 Asteraceae whole plastid genomes and concatenated nrDNA sequences were used for phylogenetic analysis. *Lactuca raddeana* and *Ainsliaea latifolia* were used as the outgroup species (Supplementary Table 2). Each region was first aligned using MUSCLE v3.8 (Edgar, 2004) and then concatenated to form six matrices, namely 1) a dataset of concatenated nrDNA sequences (aligned length 5,855 bp), 2) a dataset of 73 conserved protein-coding sequences (aligned length 61,413 bp), 3) a dataset of 95 common genes including rRNA and tRNA genes (aligned length 83,547 bp), 4) a dataset of 88 intergenic spacer regions (IGS, aligned length 37,809 bp), 5) a dataset of 73 protein sequences (aligned length 20,393 bp), and 6) the dataset of the whole plastid genomes (aligned length 121,356 bp). The maximum likelihood (ML) phylogenetic trees of the six matrices were constructed using RAxML v8.2.12 (Stamatakis, 2014) with 1000 bootstrap replicates. The GTRGAMMA substitution model was applied to the protein-coding genes, genes, IGS, and whole plastid genomes, while PROGAMMAAUTO was applied to protein sequences. The parameters for this analysis included "raxmlHPC-PTHREADS-SSE3 -f a -N 1000 -m GTRGAMMA -x 551314260 -p 551314260 -T 20". Tree visualization was performed using MEGA X (Kumar et al., 2018). Finally, the topologies recovered after analysis of the plastid and nrDNA data were compared using the dendextend package (Galili, 2015).

The analysis of phylogenetic networks was carried out using PhyloNet v.3.8.21 (Wen et al., 2018) with the command 'InferNetwork_MPL' using gene tree topologies estimated by IQ-TREE2 (Minh et al., 2020); subsequently, the phylogenetic networks in the form of Rich Newick strings were visualized in Dendroscope3 (Huson and Scornavacca, 2012). The D-statistic (ABBA–BABA) method in Dsuite (Malinsky et al., 2021) was used to test for introgression events using SNP data on *Atractylodes* plastid genomes.

Results

The structure of the *Atractylodes* plastid genome and concatenated nrDNA sequences

The plastid genome length of *Atractylodes* ranged from 152,138 bp (*A. chinensis*) to 153,268 bp (*A. lancea*). The greatest length variations among individuals of *A. chinensis*, *A. koreana*, and *A. lancea* were 956 bp, 983 bp, and 1000 bp, respectively. Each plastid genome displayed a typical quadripartite structure, consisting of a large single-copy (LSC) region (83,206~84,293 bp), a small single-copy (SSC) region (18,604~18,698 bp), and a pair of inverted repeat (IR) regions (IRa and IRb) (25,137~25,185 bp). The GC content varied from 37.69% to 37.77% and was higher in the IR regions (about 43%) than in the LSC and SSC regions (about 35% and 31%) in all species (Figure 1, Supplementary Table 3). Moreover, several long-term indels were verified by genome mapping, including the 989 bp deletion of the plastid *ndhC_trnV-UAC* IGS in *A. lancea* (HPAB0031), *A. chinensis* (HPAB0001, HPAB0003, and HPAB0006), and *A. koreana* (HPAB0010). Length differences in coding genes such as *rpoB* and *ycf2* were mainly due to the occurrence of repeat units in the sequence. A 6 bp insertion unit (TTAACC) of *rpoB* was found in *A. lancea* (HPAB0027), while a 9 bp deletion unit of *ycf2* was present in *A. chinensis* (HPAB0001).

A total of 113 unique genes were annotated in the plastid genomes, consisting of 80 protein-coding genes, 29 tRNA genes, and four rRNA genes (*rrn23S*, *rrn16S*, *rrn5S*, and *rrn4.5S*). Of these, 82 genes (61 protein-coding genes and 21 tRNA genes) were located in the LSC region. The SSC region contained 12 protein-coding genes and one tRNA gene (*trnL-UAG*).

10.3389/fpls.2022.1045423

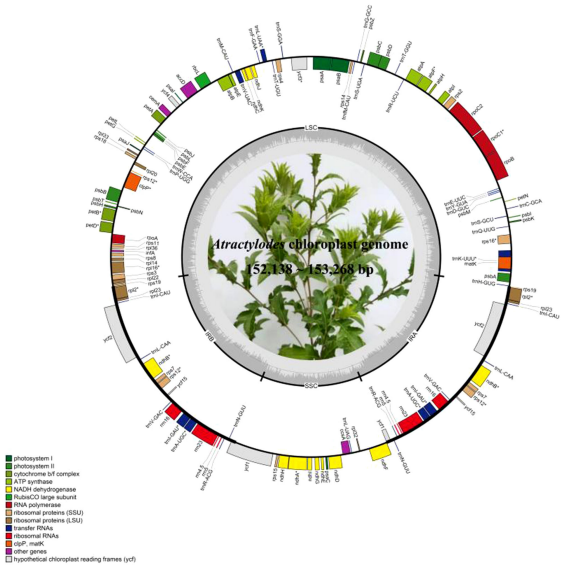


FIGURE 1
Plastid genome map of *Atractylodes* species. Gene locations outside the outer rim are transcribed in a counterclockwise direction, while genes inside are transcribed in a clockwise direction. The colored bars indicate known functional family differences. The dashed gray area in the inner circle shows the proportional GC content of the corresponding genes. LSC, large single-copy region; SSC, small single-copy region; IR, inverted repeat region.

Furthermore, 17 genes contained introns: 14 (nine protein-coding and five tRNA genes) contained one intron, and three (*rps12*, *ycf3*, and *clpP*) contained two introns (Supplementary Table 4). Small exons were also annotated in the *petB*, *petD*, and *rpl16* genes, with lengths of 6 bp, 8 bp, and 9 bp, respectively. Finally, *rps12* was identified as a trans-splicing gene.

The length of the concatenated nrDNA sequence was 5,849 bp; this sequence consisted of five parts. The lengths of 18S, ITS1, 5.8S, ITS2, and 28S were 1,809 bp, 259 bp, 158 bp, 229 bp, and 3,394 bp, respectively. The GC content varied between 55.43% and 55.62% and was higher in the ITS regions (about 63%, including ITS1 and ITS2) than in the 18S and 28S regions (about 49% and 57%) in all samples (Supplementary Table 5).

Codon usage in the *Atractylodes* plastid genomes

The amino acid frequency, codon usage, and relative synonymous codon usage (RSCU) of 80 protein-coding regions in all *Atractylodes* species were analyzed using Codon W. RSCU values ranged from 63.98 to 64.05; the number of codons ranged from 22,846 (HPAB0022) to 22,873 (HPAB0023) in 26 species; and the number of amino acids ranged from 21,706 (HPAB0023) to 22,398 (HPAB0016). Of these codons, leucine (2,275 ~ 2,338 codons) was the most abundant amino acid, with a frequency of 9.95 ~ 10.22%, while the proportion of cysteine (471 ~ 486 codons) was 2.06 ~ 2.13%; AGA (encoding arginase) and CGC (encoding arginase) were the most and least used codons, respectively (Supplementary Table 6). Almost all the amino acids had more than one synonymous codon; the exceptions were methionine and tryptophan. Furthermore, 31 codons displayed RSCU values exceeding 1. Most of the biased codons were used with A or T bases as the third codons. ATG and TGG, encoding methionine and tryptophan, exhibited no bias (RSCU = 1.00) (Supplementary Table 6).

Three types of starting codons were detected in 80 protein-coding genes. Of these, 77 genes used ATG as start codons, while two (*ndhD*, *psbL*) used ACG and one (*rps19*) used GTG. TAA, TAG, and TGA were present as stop codons in these genes. The most used stop codon was TAA at 55.56%, followed by TAG (25.92%) and TGA (18.52%). *ndhF* was the only gene that used both TAA and TGA as stop codons; of this gene, eight samples (HPAB0003, HPAB0006, HPAB0009, HPAB0010, HPAB0011, HPAB0016, HPAB0018, and HPAB0031) showed a preference for TAA and 18 utilized TGA.

The SSR units and repeat structures of *Atractylodes* plastid genomes

SSRs were detected in 24 *Atractylodes* plastid genome sequences (Supplementary Table 7). Specifically, 30, 30, 29, 30, and 30 SSR loci were detected in the *Atractylodes chinensis*, *A. koreana*, *A. lancea*, *A. japonica*, and *A. macrocephala* plastid genomes, respectively. A large proportion of the SSRs were

distributed in the LSC region (85.23%), with 12 in the SSC region and 10 in the IR regions. Polyadenine (poly-A) (34.70%, 10~16) and polythymine (poly-T) (60.41%, 10~21) represented the dominant repeats.

In addition to the SSRs, 33, 34, 36, 31, and 32 tandem repeat sequences were detected in the *A. chinensis*, *A. koreana*, *A. lancea*, *A. japonica*, and *A. macrocephala* plastid genomes, respectively. The tandem repeat lengths were 9~45 bp, and most were located in the IGS regions of the genomes. In addition, 1, 3, 5, 5, 5, 47, and 25 tandem repeats were found in the *atpI*, *ndhF*, *rpoC2*, *rps18*, *petD*, *ycf1*, and *ycf2* coding regions, respectively (Figure 2, Supplementary Table 8). Finally, an average of 42 repeat structures were revealed for each species, including F, P, R, and C repeats. P was the most common repeat type, accounting for 48.8~53.2% of all repeats, followed by F (41.9~51.2%), C (6.9%), and R (2.3%) (Figure 2, Supplementary Table 9).

Variation in *Atractylodes* plastid genome sequences and concatenated nrDNA sequences

The 24 plastid genome sequences were compared to identify differences using the online software platform mVISTA, with the *A. chinensis* plastome as the reference genome (Figure 3, Supplementary Figure 4). Highly conservative regions were evident across most of the plastid genomes, and only three relatively high-variability regions were identified in the whole genome (Figure 3). Subsequently, Mauve was used to identify local collinear blocks (LCBs) of *Atractylodes* plastid genomes (Figure 4, Supplementary Figure 5); the *A. chinensis* genome is shown at the top as the reference genome. These species showed a consistent sequence order in all the genes. The collinear blocks of all the plastid genomes, including the LSC, SSC, and IR regions, revealed relatively high levels of conservation, with no gene rearrangement.

The nrDNA sequences exhibited 53 variation sites and 25 parsimony-informative sites, accounting for 0.91% and 0.43% of the total nrDNA sequences, respectively. The majority of the variation sites were located in the ITS1 and ITS2 regions. The average inter-specific distance, average theta prime, and smallest inter-specific distance were used to characterize the inter-specific divergence, taking values of 0.0027, 0.0027, and 0.0014, respectively. The intra-specific variation was determined according to the average intra-specific difference, theta, and average coalescent depth, which yielded values of 0.0003, 0.0008, and 0.0012, respectively. DNA barcoding gaps were clearly present for species *A. carlinoides*, *A. macrocephala*, *A. japonica*, and *A. lancea*, whereas the relationship between *A. chinensis* and *A. koreana* could not be resolved.

Contraction and expansion of the IR region in *Atractylodes* plastid genome sequences

The circular structure of the plastid genome was highly conserved and generated four boundaries in the IR, LSC, and SSC regions. As the genome evolved, the contraction and expansion of the IR boundary produced different plastid genome sizes and altered certain gene locations. The *rps19* gene crossed the LSC and IRb regions in all the species. Although the *ycf1* gene was distributed across the SSC and IRb regions, most of the genes were located in the SSC region, with minor differences in length. Four indels were identified *via*

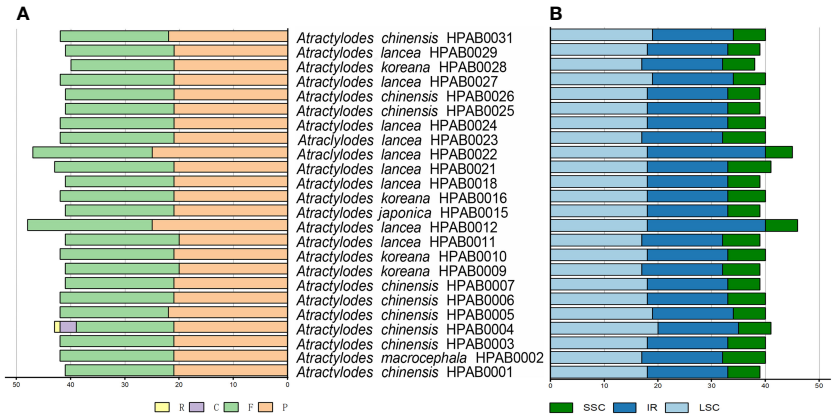


FIGURE 2
The types and distribution of repeat sequences in 24 *Atractylodes* plastid genomes. **(A)** The number of each of the four repeat types (F, forward; P, palindrome; R, reverse; C, complement). **(B)** The distribution of repeat sequences across three regions: LSC, SSC, and IR.

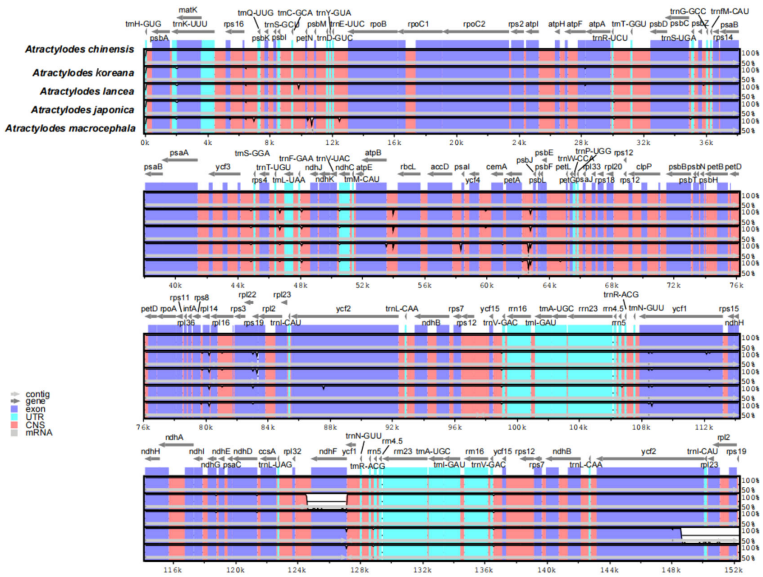


FIGURE 3
A comparison of the five plastid genomes, with *A. chinensis* as a reference, using the mVISTA alignment program. The grey arrows above the alignments show the orientation of the genes. The violet blocks indicate exons, the cyan blocks denote introns, and the salmon blocks signify conserved non-coding sequences (CNS). The y-axis indicates the identity percentage, ranging from 50% to 100%, while the x-axis represents sequence length.

multiple sequence alignment in the *ycf1* gene. Specifically, a TTTGAA insertion was detected in positions 4,435-4,440 in *A. chinensis* and *A. macrocephala*; an AAGACGAA deletion occurred at positions 800-808 in *A. chinensis*; an AAATAC deletion was evident at positions 4,290-4,295 in *A. lancea*; and finally, an AAGACGAAG insertion was detected at positions 791-799 in *A. macrocephala*. The *ndhF* gene and the *ycf1* pseudogene were detected at the junction of SSC and IRa. The *ndhF* gene was mainly located in the SSC region but spanned the junction 15 bp into the IRB region. The *ycf1* pseudogene was entirely located in the IRa region. Additionally, the *rps19* pseudogene and *trnH* gene were located at the junction of the IRa and LSC regions, while *rps19* spanned the junction 1 bp into the LSC region (Figure 5, Supplementary Figure 6).

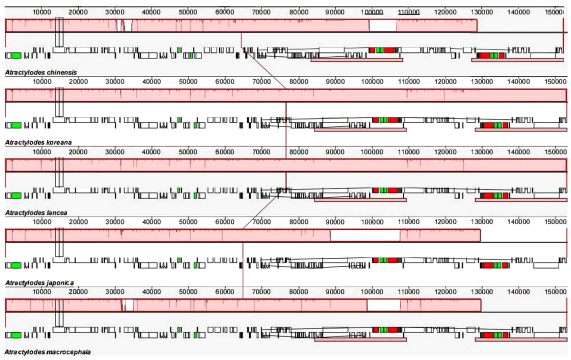


FIGURE 4
A comparison of the whole plastid genomes of the *Atractylodes* species using the Mauve algorithm. The red LCBs indicate syntenic regions, while the histograms within each block represent the degree of sequence similarity. rRNA, protein-coding, and tRNA gene annotations are denoted by red, white, and green boxes, respectively.

Heteroplasmic variations in the whole plastid genome

A total of 64 heteroplasmic variations were detected in the whole plastid genome of *Atractylodes* (Supplementary Table 10). This heteroplasmy consisted primarily of two types, namely heteroplasmy in insertion/deletions (indels) and heteroplasmy in single nucleotide polymorphism sites (SNPs). For example, an *ndhF* gene sequence deletion of 20 bp in length was identified in eight samples (HPAB0003, HPAB0006, HPAB0009, HPAB0010, HPAB0011, HPAB0016, HPAB0018, and HPAB0031) *via* multiple sequence alignment. Further genome mapping simultaneously detected two types of reads (insertion and deletion) in the intra-plastid genome, such as in sample HPAB0016. A further example is that the base R representing adenine (A) and guanine (G) was found in the assembly result for HPAB0001. Subsequently, a total of 235 reads related to this region were extracted from the shotgun sequencing data,

indicating the simultaneous detection of G (185) and A (50) at the corresponding site (Supplementary Table 10).

Phylogenetic tree, phylogenetic network, and gene introgression analyses

The phylogenetic relationships of *Atractylodes* were analyzed using six data sets of concatenated nrDNA sequences, 73 conserved plastid protein-coding gene sequences, 95 plastid common gene sequences, 88 plastid IGS regions, 73 plastid protein sequences, and the whole plastid genome sequences (Supplementary Table 11). The resulting phylogenetic trees showed that *Atractylodes* was a monophyletic clade related to *Tugarinovia mongolica*. The concatenated nrDNA sequences and plastid phylogenetic analyses indicated that *A. carlinoides* separated from the rest of the *Atractylodes* species with a high bootstrap value. *A. macrocephala* alone formed a relatively independent clade, with the *A. lancea* complex as a sister group of this.

The phylogenetic tree generated for the concatenated nrDNA sequence dataset showed that the *A. lancea* complex was divided into three subclades: *A. lancea*, *A. japonica*, and *A. chinensis–A. koreana*. The *A. lancea* clade included ten samples (HPAB0011, HPAB0012, HPAB0018, HPAB0021, HPAB0022, HPAB0023, HPAB0026, HPAB0028, HPAB0031, and MG874804). The *A. japonica* clade included three samples (HPAB0015, MW301112, and MT834523), which were outside the *A. chinensis–A. koreana* clade with a bootstrap value of 91. The most controversial aspect of the tree was the *A. chinensis–A. koreana* clade, which included 17 A. *chinensis* and *A. koreana* samples (Figure 6, Supplementary Figure 7).

The phylogenetic trees for the five different plastid datasets presented similar topologies (Supplementary Figures 8-Supplementary Figure 12). The whole plastid genome dataset yielded better-supported trees than the other four datasets. In this tree, the *A. lancea* complex was divided into one small and one large clade. The small clade was a sister of *A. macrocephala* and contained two *A. chinensis* samples (HPAB0003 and HPAB0006), one *A. lancea* sample (HPAB0031), and two *A. koreana* samples (HPAB0010 and HPAB0016) with a bootstrap value of 100. The remaining 25 samples, consisting of nine *A. lancea*, three *A. japonica*, four *A. koreana*, and nine *A. chinensis*, formed a large clade with a bootstrap value of 55. The phylogeny of the four species in this large clade was ambiguous and could not be clearly resolved (Figure 6). Phylogenetic analyses for the nrDNA and whole plastid genome indicated plastid and nuclear discordance.

PhyloNet was used to further assess putative hybridization events in the phylogeny. The analysis indicated that certain loci in the genome of *Atractylodes* shared a most recent common ancestor with loci in that of *Arctium lappa*, and others shared a most recent common ancestor with loci in that of *Tugarinovia mongolica*; this was the case in all networks allowing 1-4 reticulations (Figure 7). When 4 reticulations were allowed, the resulting network showed that *A. lancea* may have hybridized with the common ancestor of *A. chinensis* and *A. japonica*. The results of ABBA–BABA tests showed that *A. chinensis* was more closely related to *A. japonica* than to *A. lancea* (Supplementary Table 12).

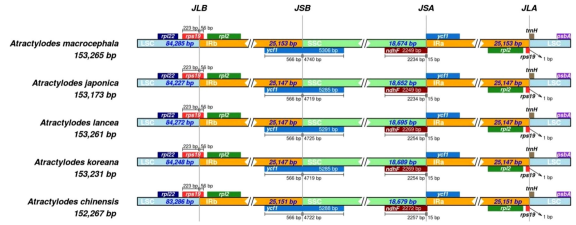


FIGURE 5
A comparison of the LSC and IRb border region and the SSC and IRa border region for the five *Atractylodes* species. JLB: junction of LSC and IRb; JSA: junction of SSC and IRa. JSB: junction of IRb and SSC; JLA: junction of IRa and LSC.

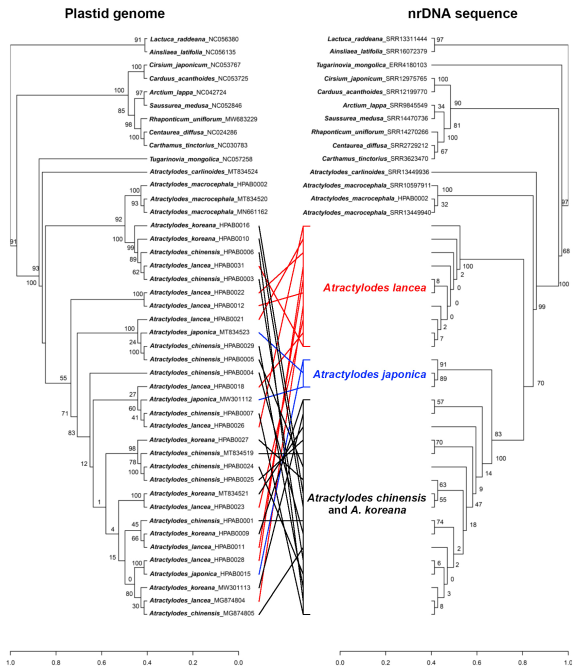


FIGURE 6
A comparison of the topologies recovered *via* plastid data (left) and *via* concatenated nrDNA sequences (right). *A. lancea* are shown in red, *A. japonica* in blue, and the *A. chinensis–A. koreana* clade in black.

Discussion

Atractylodes is a small genus of the Asteraceae family mainly distributed across East Asia. It is a cross-pollination plant group, meaning that its morphological character is extremely susceptible to environmental factors (Hu et al., 2000). The phylogenetic relationships and the taxonomic treatment within the entire genus are intricate due to the continuous nature of variation in its morphological features (Peng et al., 2012). Atractylodes rhizomes distributed at low altitudes are bead-shaped and run horizontally; however, with changes in altitude and the ecological environment, rhizomes at medium altitudes appear in clumps and grow obliquely downwards (Zhu, 1981). The petioles and degree of leaf-splitting of the genus are still undergoing continuous evolution (Peng and Wang, 2007). In some instances, the lower and middle cauline leaves are petiolate, whereas basal leaves are sometimes subsessile (Hu et al., 2000). Although the leaf blades are generally divided into 3-5 pinnately arranged segments, they are occasionally undivided near the base, with a few small spiny lobes (Zhu, 1981). The 24 Atractylodes samples collected in this study also exhibited several continuously varying morphological features, especially those samples cultivated in Hubei province.

Consistent with previous studies, the phylogenetic trees presented here indicate that A. carlinoides is a basal species in the Atractylodes genus, and is a sister of the remainder of the Atractylodes species. Unlike one previous study predicting that A. macrocephala may be an A. chinensis and A. lancea hybrid (Wang et al., 2020), this study found that A. macrocephala forms an independent branch, with a bootstrap value of 100 in the case of both plastid data and nrDNA data. Regarding A. koreana, this species is mainly distributed in the Liaoning and Shandong provinces of China. Its lower and middle cauline leaves are undivided, which represents a point of distinction from the morphological characteristics of the other Atractylodes species. However, the ITS genotype of A. koreana was found here to be consistent with that of A. chinensis, and this also has been reported in a previous study (Shiba et al., 2006). A. chinensis and A. lancea are generally discriminated on the basis of differences in the shapes of their leaves. However, Shi has indicated that these morphological differences are unstable (Zhu, 1981), and past authors have been misled because the number of specimens they possessed was extremely limited, resulting in a poor understanding of the polytype of this species. Here, ten samples of A. lancea formed an independent clade in an analysis using nrDNA data from both wild and cultivated samples. A. japonica has long petioles, and the leaf blades generally divide almost to the base into 3-5 segments; these traits constitute obvious differences from other species in the A. lancea complex. This label has been considered a synonym for A. lancea, as recorded in the latest version of Flora of China. However, this study identified multiple strands of supporting evidence for a close relationship between A. japonica and other variations of the A. lancea complex in the nrDNA concatenated sequence-based phylogeny. Moreover, species trees obtained using PhyloNet confirmed that A. japonica is closely related to A. chinensis but separate from A. lancea.

Cytonuclear discordance can be caused by many factors, such as ancient hybridization and gene introgression. It is a common phenomenon in plant systematics and has been reported in many genera, such as section *Galoglychia* (Renoult et al., 2009) and *Cotoneaster* (Meng et al., 2021). Existing

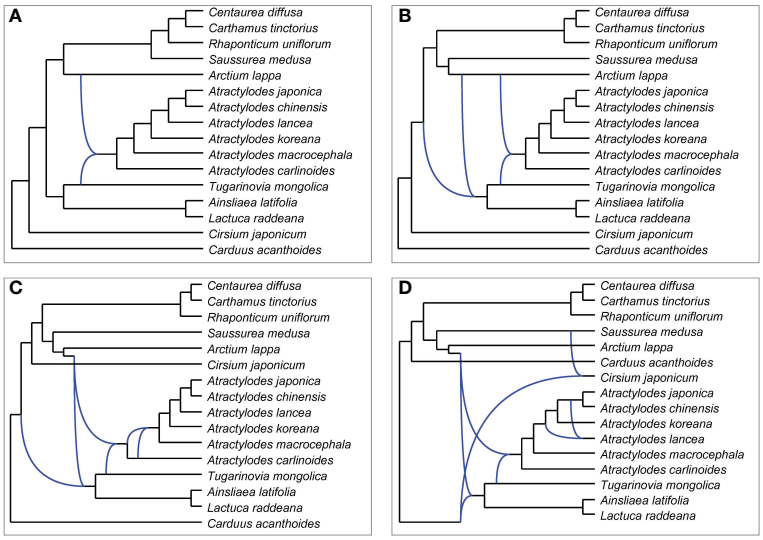


FIGURE 7
Networks representing plastid gene trees generated by PhyloNet MPL, allowing a maximum of 1 **(A)**, 2 **(B)**, 3 **(C)**, or 4 **(D)** reticulations.

cpDNA and nuclear genetic evidence has revealed that sect. *Galoglychia* has many obvious nuclear–cytoplasmic phylogenetic tree conflicts, which are likely to be caused by ancient hybridization followed by gene introgression. In the case of *Cotoneaster*, sequences of the complete plastid genomes and 204 low-copy nuclear genes of 69 species were used for the phylogenetic analysis, and the results revealed there were conflicts between the plastid genome and low-copy nuclear phylogenies at both the species and clade levels. These instances of cytonuclear discordance may be caused by frequent hybridization events and incomplete lineage sorting (ILS). For species of *Atractylodes*, gene introgression usually results from natural hybridization among closely related species in sympatric populations (Hu et al., 2000). Shiba et al. have indicated that the continuous morphological variation in features between *A. lancea* and *A. chinensis* may be caused by the presence of such hybrids. Interspecific hybridization between *A. lancea* and *A. chinensis* has been observed in 25 samples containing nucleotide additives (Peng and Wang, 2007). Here, the phylogenetic network analysis of plastid genes revealed that *A. lancea* may have hybridized with the common ancestor of *A. chinensis* and *A. japonica*.

In conclusion, the results of this study tended to support treatment of *A. japonica* as an independent species. Although samples of *A. lancea* formed an independent clade, the other species in the *A. lancea* complex were still mingled with one another due to a complicated pattern of evolution in this genus, as shown by the phylogeny according to the plastid genomic data. In addition, this study revealed extensive discordance based on the cytonuclear genomic data, primarily involving the *A. lancea* complex. In future research, analysis of the A. lancea complex with sufficient single-copy nuclear genes or use of a reduced-representation genomic approach will be necessary to clarify the genetic differentiation.

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

LS, JL, and XiaoZ conceived and designed the study. HX, CZ, XinZ, QW, and YL collected and identified the plant materials. JL, MS, and HX performed the experiments. JL, MS, ZZ, and WK analyzed the data. JL, LS, MS, and HX wrote the manuscript. JL, LS, and XiaoZ revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by National Key R&D Program: Intergovernmental Cooperation in International Science and Technology Innovation (2022YFE0119300), China Postdoctoral Science Foundation (2022M720504), Beijing Municipal Natural Science Foundation (7202136), the National Natural Science Foundation of China (81703659), Guangxi Science and Technology base and talent project (AD22080012).

Acknowledgments

We thank Dr Ran Wei, State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, for providing help with the phylogenic analysis conducted in this study.

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.2022.1045423/full#supplementary-material

References

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

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

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, e8613. doi: 10.1371/journal.pone.0008613

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

Deng, A.-P., Wu, Z.-T., Liu, T., Kang, L.-P., Nan, T.-G., Zhan, Z.-L., et al. (2016). Advances in studies on chemical compositions of atractylodes lancea and their biological activities. *Zhongguo Zhong Yao Za Zhi* 41, 3904–3913. doi: 10.4268/cjcmm20162104

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinf.* 5, 113. doi: 10.1186/1471-2105-5-113

Galili, T. (2015). Dendextend: an r package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* 31, 3718–3720. doi: 10.1093/bioinformatics/btv428

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

Hu, S., Feng, X., Ji, L., and Nie, S. (2000). Atractylodes lancea and its geo-varieties. *Chin. Traditional Herbal Drugs* 31, 781–784. Available at: https://www.tiprpress.com/zcy/article/abstract/20001033?st=search

Huson, D. H., and Scornavacca, C. (2012). Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Systematic Biol.* 61, 1061–1067. doi: 10.1093/sysbio/sys062

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

Kim, J.-H., Doh, E.-J., and Lee, G. (2016). Evaluation of medicinal categorization of atractylodes japonica koidz. by using internal transcribed spacer sequencing analysis and HPLC fingerprinting combined with statistical tools. *Evidence-Based Complementary Altern. Med.* 2016, 2926819. doi: 10.1155/2016/2926819

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

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

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

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

Lin, R., and Shi, Z. (1987). *Flora reipublicae popularis sinicae* (Beijing: Compositae Science Press).

Liu, S. (1959). *Northeast medicinal flora* (Beijing: Science Press).

Liu, H., He, J., Ding, C., Lyu, R., Pei, L., Cheng, J., et al. (2018). Comparative analysis of complete chloroplast genomes of anemoclema, anemone, pulsatilla, and hepatica revealing structural variations among genera in tribe anemoneae (Ranunculaceae). *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01097

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

Malinsky, M., Matschiner, M., and Svardal, H. (2021). Dsuite-fast d-statistics and related admixture evidence from VCF files. *Mol. Ecol. Resour.* 21, 584–595. doi: 10.1111/1755-0998.13265

McPherson, H., van der Merwe, M., Delaney, S. K., Edwards, M. A., Henry, R. J., McIntosh, E., et al. (2013). Capturing chloroplast variation for molecular ecology studies: a simple next generation sequencing approach applied to a rainforest tree. *BMC Ecol.* 13, 8. doi: 10.1186/1472-6785-13-8

Meleshko, O., Martin, M. D., Korneliussen, T. S., Schröck, C., Lamkowski, P., Schmutz, J., et al. (2021). Extensive genome-wide phylogenetic discordance is due to incomplete lineage sorting and not ongoing introgression in a rapidly radiated bryophyte genus. *Mol. Biol. Evol.* 38, 2750–2766. doi: 10.1093/molbev/msab063

Meng, K.-K., Chen, S.-F., Xu, K.-W., Zhou, R.-C., Li, M.-W., Dhamala, M. K., et al. (2021). Phylogenomic analyses based on genome-skimming data reveal cyto-nuclear discordance in the evolutionary history of cotoneaster (Rosaceae). *Mol. Phylogenet. Evol.* 158, 107083. doi: 10.1016/j.ympev.2021.107083

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., et al. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37, 1530–1534. doi: 10.1093/molbev/msaa015

Peng, H. S., and Wang, D. Q. (2007). Studies on population biology of transitional types of genus atractylodes in anhui province. *Zhongguo Zhong Yao Za Zhi* 32, 793–797. Available at: http://www.cjcmm.com.cn/WKE3/WebPublication/paperDigest.aspx?paperID=5a670764-5069-4d94-a816-9b844c4d0625

Peng, H.-S., Yuan, Q.-J., Li, Q.-Q., and Huang, L.-Q. (2012). Molecular systematics of genus atractylodes (Compositae, cardueae): Evidence from internal transcribed spacer (ITS) and trnL-f sequences. *Int. J. Mol. Sci.* 13, 14623–14633. doi: 10.3390/ijms131114623

Renoult, J. P., Kjellberg, F., Grout, C., Santoni, S., and Khadari, B. (2009). Cyto-nuclear discordance in the phylogeny of ficus section galoglychia and host shifts in plant-pollinator associations. *BMC Evolutionary Biol.* 9, 248. doi: 10.1186/1471-2148-9-248

Scarcelli, N., Mariac, C., Couvreur, T. L. P., Faye, A., Richard, D., Sabot, F., et al. (2016). Intra-individual polymorphism in chloroplasts from NGS data: where does it come from and how to handle it? *Mol. Ecol. Resour.* 16, 434–445. doi: 10.1111/1755-0998.12462

Shi, Z. (2011). *Flora of China volume 20– 21 (Asteraceae)* (Beijing: Science press).

Shiba, M., Kondo, K., Miki, E., Yamaji, H., Morota, T., Terabayashi, S., et al. (2006). Identification of medicinal atractylodes based on ITS sequences of nrDNA. *Biol. Pharm. Bull.* 29, 315–320. doi: 10.1248/bpb.29.315

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, 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

Sun, S.-S., Zhou, X.-J., Li, Z.-Z., Song, H.-Y., Long, Z.-C., and Fu, P.-C. (2019). Intra-individual heteroplasmy in the gentiana tongolensis plastid genome (Gentianaceae). *PeerJ* 7, e8025. doi: 10.7717/peerj.8025

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, L., Zhang, H., Wu, X., Wang, Z., Fang, W., Jiang, M., et al. (2020). Phylogenetic relationships of atractylodes lancea, a. chinensis and a. macrocephala, revealed by complete plastome and nuclear gene sequences. *PloS One* 15, e0227610. doi: 10.1371/journal.pone.0227610

Wen, D., Yu, Y., Zhu, J., and Nakhleh, L. (2018). Inferring phylogenetic networks using PhyloNet. *Systematic Biol.* 67, 735–740. doi: 10.1093/sysbio/syy015

Yin, M., Xiao, C.-C., Chen, Y., Wang, M., Guan, F.-Q., Wang, Q.-z., et al. (2015). A new sesquiterpenoid glycoside from rhizomes of atractylodes lancea. *Chin. Herbal Medicines* 7, 371–374. doi: 10.1016/S1674-6384(15)60066-1

Zhu, S. (1981). On the nomenclature of Chinese drug "Cangzhu". *Acta Phytotaxonomica Sin.* 19, 318–322. Available at: https://www.plantsystematics.com/CN/Y1981/V19/I3/318

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 20 December 2022
DOI 10.3389/fpls.2022.1061174

OPEN ACCESS

EDITED BY
Wenpan Dong,
Beijing Forestry University, China

REVIEWED BY
Zhiqiang Wu,
Chinese Academy of Agricultural Sciences, China
Xiao-Jian Qu,
Shandong Normal University, China

*CORRESPONDENCE
Shu-Miaw Chaw
smchaw@sinica.edu.tw

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

RECEIVED 04 October 2022
ACCEPTED 23 November 2022
PUBLISHED 20 December 2022

CITATION
Wu C-S, Chen C-I and Chaw S-M (2022) Plastid phylogenomics and plastome evolution in the morning glory family (Convolvulaceae). *Front. Plant Sci.* 13:1061174. doi: 10.3389/fpls.2022.1061174

Plastid phylogenomics and plastome evolution in the morning glory family (Convolvulaceae)

Chung-Shien Wu[1], Chung-I. Chen[2] and Shu-Miaw Chaw[1]*

[1]Biodiversity Research Center, Academia Sinica, Taipei, Taiwan, [2]Department of Forestry, National Pingtung University of Science and Technology, Pingtung, Taiwan

Convolvulaceae, the morning glories or bindweeds, is a large family containing species of economic value, including crops, traditional medicines, ornamentals, and vegetables. However, not only are the phylogenetic relationships within this group still debated at the intertribal and intergeneric levels, but also plastid genome (plastome) complexity within Convolvulaceae is not well surveyed. We gathered 78 plastomes representing 17 genera across nine of the 12 Convolvulaceae tribes. Our plastid phylogenomic trees confirm the monophyly of Convolvulaceae, place the genus *Jacquemontia* within the subfamily Dicranostyloideae, and suggest that the tribe Merremieae is paraphyletic. In contrast, positions of the two genera *Cuscuta* and *Erycibe* are uncertain as the bootstrap support of the branches leading to them is moderate to weak. We show that nucleotide substitution rates are extremely variable among Convolvulaceae taxa and likely responsible for the topological uncertainty. Numerous plastomic rearrangements are detected in Convolvulaceae, including inversions, duplications, contraction and expansion of inverted repeats (IRs), and losses of genes and introns. Moreover, integrated foreign DNA of mitochondrial origin was found in the *Jacquemontia* plastome, adding a rare example of gene transfer from mitochondria to plastids in angiosperms. In the IR of *Dichondra*, we discovered an extra copy of *rpl16* containing a direct repeat of ca. 200 bp long. This repeat was experimentally demonstrated to trigger effective homologous recombination, resulting in the coexistence of intron-containing and -lacking *rpl16* duplicates. Therefore, we propose a hypothetical model to interpret intron loss accompanied by invasion of direct repeats at appropriate positions. Our model complements the intron loss model driven by retroprocessing when genes have lost introns but contain abundant RNA editing sites adjacent to former splicing sites.

KEYWORDS

convolvulaceae, plastome, phylogenomics, gene transfer, intron loss

Introduction

Convolvulaceae, commonly known as the morning glories or bindweeds, is a large family sister to Solanaceae (Stefanović et al., 2002) in the order Solanales (AGP IV, 2016). Morning glories comprise approximately 2,000 species in 60 genera (Simões et al., 2022), with a diverse range of morphological characteristics from herbs, shrubs, vines, to parasites *Cuscuta* (dodders). They mostly inhabit tropical areas and include many economically valuable crop (e.g., sweet potato), traditional medicine (e.g., dodders), ornamental (e.g., morning glory), and vegetable (e.g., water spinach) species. Using a few plastid and nuclear markers, early molecular phylogenetic studies classified 12 tribes in Convolvulaceae (Stefanović et al., 2002; Stefanovi´c et al., 2003), but controversies remain at the intertribal and intergeneric levels even when organellar (Lin et al., 2022) and nuclear (Simões et al., 2022) genomic data were used for comparative phylogenetic analyses.

Plastids are cellular organelles and have their own genomes, called plastomes. In general, seed plant plastomes are structurally conserved and contain two inverted repeats (IRs) separating the sequences into a large single copy (LSC) region and a small single copy (SSC) one. Integration of foreign DNA into plastomes was initially thought impossible due to the absence of active DNA import systems in plastids (Smith, 2011). However, plastid mitochondrial (mt) DNA-derived sequences (PTMTs) were later discovered in a handful of species (e.g., Iorizzo et al., 2012a; Iorizzo et al., 2012b; Straub et al., 2013; Ma et al., 2015; Rabah et al., 2017; Raman et al., 2019; Raman et al., 2021), suggesting that acquisition of foreign DNA that shapes plastome complexity is possible, although the mechanism is unclear. Plastomes also undergo rearrangements, including inversions and changes of gene content through deletions and duplications (see review in Wicke et al., 2011). Gene content changes due to IR contraction and expansion frequently occurred during plant evolution (Wang et al., 2008; Zhu et al., 2016). A comparison of plastid transcriptomes has shed light on the effects of inversions on gene expression in conifer plastids (Wu et al., 2021). However, to date there is little conclusive evidence as to the consequences of plastid gene content changes.

Plastids contain several group II introns. Unlike their eubacterial counterparts/ancestors, modern plastid group II introns have lost their mobility and rely on host-encoded factors for splicing (Lambowitz and Zimmerly, 2011). Group II introns are generally 400–800 bp long and possess six major domains (I–VI domains) that form a conserved tertiary structure and an active site for splicing (Kelchner, 2002; Lambowitz and Zimmerly, 2011). Precise deletions of introns have been proposed to be a consequence of retroprocessing operated in plastids (Downie et al., 1991), while localized retroprocessing appears to be a widespread mechanism underlying intron loss in plant mitochondria (Cuenca et al., 2016).

Recent studies of autotrophic plastomes in Convolvulaceae were mainly confined to the genus *Ipomoea* (Eserman et al., 2014; Park et al., 2018; Sun et al., 2019; Laux et al., 2022). However, Lin et al. (2022) compared plastomes across eight Convolvulaceous tribes and found several unusual features, such as atypical IRs, gene and intron losses, and insertions of foreign DNA of unknown origin. These studies suggest that Convolvulaceous plastomes are labile and their complexity might be underestimated. To further explore this variation, we gathered 78 plastomes representing 17 genera from nine of the 12 tribes in the morning-glory family, including *Jacquemontia*, considered to have the greatest sequence divergence among autotrophic Convolvulaceae (Stefanović et al., 2002). These data were used to construct phylogenetic trees, estimate nucleotide substitution rates, and characterize plastomic rearrangements. Furthermore, we report the first PTMT case in Convolvulaceae and propose a repeat-mediated model to interpret intron loss without retroprocessing.

Materials and methods

Taxon sampling, DNA and RNA extraction, and sequencing

We collected 19 Convolvulaceae taxa. Their voucher information is shown in Table S1. Total DNA was extracted from fresh leaves using the CTAB method described in Stewart and Via (1993). The extracted DNA was used for library construction using Celero™ DNA-Seq Library Preparation Kits. Total RNA was extracted from *D. micrantha* leaves using Plant Total RNA Purification Kits (GeneMark, Taiwan). After DNase I treatment, the NuQuant® Universal Plus mRNA-Seq Kit, designed for poly (A) selection, was used to prepare cDNA libraries. We used the Illumina HiSeq 4000 platform at Genomics BioSci & Tech (New Taipei City, Taiwan) to obtain approximately 10 Gb of paired-end (PE) reads per library.

Genome assembly and annotation

Adaptors and low-quality reads were trimmed using Trimmomatic v0.36 (Bolger et al., 2014) with default parameters. Plastomes were assembled using NOVOPlasty v4.3 (Dierckxsens et al., 2017) and GetOrganelle v1.7.6.1 (Jin et al., 2020) with our own databases containing available *Ipomoea* plastomes. We merged the results generated by the two assemblers and did base-scale corrections using Pilon v1.24 (Walker et al., 2014). The assembled scaffolds were annotated in Geneious Prime (https://www.geneious.com/), with the *I. aquatica* plastome (NC056300) as the reference. Protein-coding genes and tRNAs were further confirmed by aligning them to

their orthologs and tRNAscan-SE v2.0 (Chan et al., 2021), respectively.

Sequence alignments and tree construction

Sequences of 78 plastid protein-coding genes were retrieved from our assemblies and other publicly available plastomes (Table S2). Sequence alignments were conducted using MUSCLE (Edgar, 2004) implemented in MEGA 7 (Kumar et al., 2016). Concatenations of the alignments were performed in Geneious Prime, yielding a supermatrix of 77,232 bp for downstream analyses. We used PartitionFinder v2.1.1 (Lanfear et al., 2017) to search the best scheme of data partitions under the Bayesian information criterion (BIC). This scheme was subsequently incorporated in building ML and BI trees using IQ-tree v2.2.0 (Minh et al., 2020) and MrBayes v3.2.7 (Ronquist et al., 2012), respectively. Supporting values for nodes of the ML tree were estimated from 5,000 ultrafast bootstraps. Two independent runs were conducted for 2×10^7 generations and one tree per 1,000 generations was sampled for BI tree construction. The initial 25% of the sampled trees were discarded as burin-in and Tracer v1.7.1 (Rambaut et al., 2018) was used to check if the effective sample size (ESS) exceeded 200 in all parameters.

Read mapping

DNA and RNA mapping analyses were performed in Geneious Prime, with the option of "Map to Reference" and mappers = "Geneious" for DNA-seq and "Geneious RNA" for RNA-seq reads. Gaps were not allowed during mapping.

Calculation of absolute synonymous (*RS*) and nonsynonymous (*RN*) substitution rates

To facilitate calculations, the supermatrix mentioned above was trimmed to contain only 68 accessions of unique species. This trimmed matrix was then used to estimate synonymous (*dS*) and nonsynonymous (*dN*) trees using Codeml of the Paml 4.9j package (Yang, 2007) under a branch model and codon frequency = F3 × 4. Molecular dating was conducted using the Paml MCMCTree module. The constrained ages were based on the settings described in Eserman et al. (2014) with a few modifications: the crown group for Solanaceae (23.0–33.9), for Convolvulaceae (47.8–56.0), for Ipomoeeae *s.l.* + *Merremia* (41.2–47.8 MYA). Markov chain Monte Carlo was evaluated for 2×10^6 generations under a HKY85 + G5 model. We set the sampling frequency = 10 and burnin = 25%. Terminal branches leading to species were used to calculate species *RS* and *RN* from dividing *dS* and *dN* branch lengths by the duration of species evolution.

PCR assays

We verified the existence of mtDNA insertions using PCR and two specific primer sets (Set 1: F1, 5'-ACTTCT GGTTCCGGCGAACG-3'and R1, 5'-GCTAGAATTAGC ATTATGAGCTGCTCG-3'; and Set 2: F2, 5'-CAATTCATTC GTAGTAGGATATTGAAACCTC-3'and R2, 5'-AAGCGCATA ATTGGTTGAAGATCAC-3'). Targeted fragments were amplified in a 20 µl PCR reaction: 9.6 µl water, 3.2 µl dNTP mixture (2.5 mM each), 1 µl 10 µM of each primer, 0.2 µl TaKaRa LA Taq®, 2 µl 10X LA PCR buffer II, 2 µl 25 mM $MgCl_2$, and 1 µl of genomic DNA (~25 ng). Amplification conditions were as follows: initial denaturation at 94°C for 3 min., followed by 30 cycles of 98°C for 15 sec., 58°C for 20 sec., and 68°C for 5 min. The PCR was finished with a final step at 72°C for 10.0 min. We also designed PCR primers to verify the coexistence of *rpl16*-LSC (*rpl16*-LSC-F: 5'-GAGAGTT TCTTCTCATCCAGCTCCTC-3' and *rpl16*-LSC-R: 5'-CGGAACCTGTGAATGCAAAAGATC-3'), *rpl16*-IR (*rpl16*-IR-F: 5'- GATTAGGGTAAACCAGACCCATTCATAGT-3' and *rpl16*-IR-R: 5'-ATTCTTCCTCTATGTTGTTTACG GAATCTG-3'), and *rpl16*-IR* (*rpl16*-IR*-F: 5'-CTTTT GATATAATTATCATTGCTATGCTTAGTCC-3' and *rpl16*-IR*-R: 5'-AATTGAGTTCGTATAGGCATTTTGGATG-3') copies. PCR conditions were similar to those used in examining mtDNA insertions, but the number of PCR cycles was gradually increased from 10 to 30. Electrophoresis was carried out for an hour at 50 V on 0.8% agarose TAE gels and PCR amplicons were visualized in a Quantum CX5 gel image system.

Results

Plastid phylogenomics of Convolvulaceae

The 19 newly sequenced plastomes are circular molecules. They exhibit typical quadripartite structure of a large single-copy (LSC) region, a small single-copy (SSC) region, and a pair of inverted repeats (IRs). Their sizes range from 152,365 to 165,459 bp and GC content from 37.3 to 38.8%. Using Solanaceae as the outgroup, our ML and BI trees show nearly identical topologies (Figure 1; Figure S1). Both trees suggest that the tribes Cardiochlamyeae and Erycibeae constitute the earliest diverged clade, followed by Cuscuteae, a solo parasitic tribe in Convolvulaceae. The tribes Cresseae *s.l.*, Dichondreae *s.l.*, and Jacquemontieae form a monophyletic clade sister to the clade

10.3389/fpls.2022.1061174

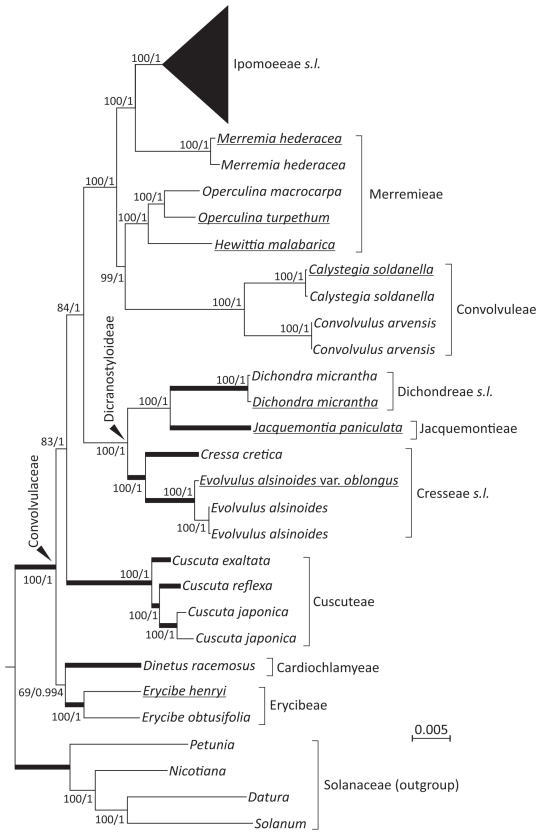


FIGURE 1
Plastid phylogenomics of representative Convolvulaceous genera and tribes. The topology is based on a ML tree inferred from a concatenation of 78 plastid protein-coding genes. Values along branches reflect bootstrap support (%) for ML trees and posterior possibilities for Bayesian inference (BI) trees. Thick branches are shortened to 1/10 of their real lengths. Underlined taxa were sequenced in this study.

comprising the tribes Convolvuleae, Ipomoeeae *s.l.*, and Merremieae. Notably, our trees strongly support (BS = 100% and PP = 1) Merremieae as a paraphyletic tribe because the position of *Merremia* is closer to Ipomoeeae *s.l.* than to other Merremieae genera (Figure 1).

Within Ipomoeeae *s.l.*, two separated clades are observed: Argyreiinae and Astripomoeinae, with the former containing the four genera *Argyreia*, *Ipomoea*, *Stictocardia*, and *Turbina* and the latter only *Ipomoea* (Figure S1). In the ML tree, the three *I. aquatica* accessions, including two cultivars, are monophyletic and sister to *I. reptans* with weak support (63%). Such relationships, however, are not present in the BI tree under the 50% majority rule (Figure S1). To date, *I. reptans* has been considered a synonym for *I. aquatica*. Here, we treat *I. reptans* as a separate species because of its semi-terrestrial habitat, purple stems, and corollaceous color differing from *I. aquatica* (Figure S1).

Extreme variability in plastid nucleotide substitution rates

To explicitly assess the variation in substitution rates across Convolvulaceae, synonymous (*dS*) and nonsynonymous (*dN*) trees were inferred from the concatenation of the 78 plastid genes. This concatenated gene set was also used to estimate divergence times among lineages. Our molecular dating suggests that Convolvulaceae split from Solanaceae at ca. 88.7 MYA and divergence of the Convolvulaceous tribes occurred in between 23.5 and 52.8 MYA (Figure S2). Figure 2 depicts absolute synonymous (*RS*) and nonsynonymous (*RN*) substitution rates that allow for rate comparisons on the same time scale. Notably, the *RS* and *RN* are strongly correlated across Convolvulaceae (R^2 = 0.7395), suggesting their dependent evolution. We noticed that Ipomoeeae *s.l.*, Merremieae, Convolvuleae, and *Erycibe* evolved at an equal rate but significantly slower than *Dichondra*, *Jacquemontia*, Cresseae *s.l.*, *Cuscuta*, and *Dinetus* in *RS* (Mann Whitney test, $P < 0.05$). In particular, the mean *RS* in Cresseae s.l. is approximately seven times faster than in Ipomoeeae s.l. Significantly elevated *RN* is only detected in Cresseae *s.l.* and *Cuscuta* (Mann Whitney test, $P < 0.05$). Taken together, these results show that substitution rates at *dS* sites are highly variable across Convolvulaceae.

Plastomic rearrangements and IR contraction/expansion

To sketch plastomic rearrangement profile across Convolvulaceae evolution, we labeled rearrangements in SC regions (hereafter called SC rearrangements) on the tree branches (Figure 3; left panel) and compared IR gene content among sampled taxa (Figure 3; right panel). SC Rearrangements in *Cuscuta* taxa are not included in the current study because they have been described previously (Braukmann et al., 2013; Banerjee and Stefanović, 2020). We discovered 13 SC rearrangements across the tree. All Convolvulaceous genera, including *Cuscuta*, lack the *rpl2* intron and *infA* gene, and the latter is also absent in Solanaceae (Amiryousefi et al., 2018). This implicates two ancient loss events that occurred before (i.e., *infA*) and after (*rpl2* intron) the split of Convolvulaceae from Solanaceae (Figure 3). In contrast, the remaining SC rearrangements are all genus specific. For instance, an inversion of the *atpI–atpB* region, duplication of *psbT*, and loss of *rpl23* in *Dinetus*; integration of mtDNA between *psaA* and *ycf3* in *Jacquemontia*; an inversion of the *trnL–atpB* region, loss of *rpl23*, and losses of *rps16* and *ycf3* introns in *Dichondra*; losses of *rps16* and *rpoC1* introns in *Evolvulus*; and loss of *rpl23* in *Operculina*. Overall, our analyses suggest that *rpl23* has been independently lost at least three times during Convolvulaceae evolution (Figure 3).

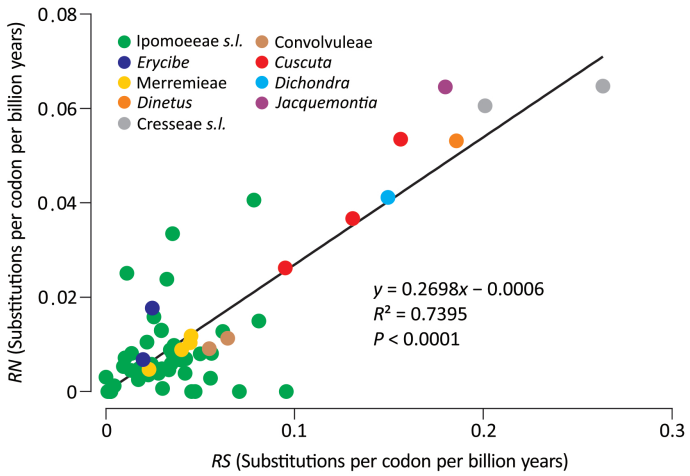


FIGURE 2
Comparisons of absolute synonymous (*RS*) versus nonsynonymous (*RN*) substitution rates across Convolvulaceae. The solid line is the regression between *RN* and *RS*.

Using Solanaceae and ancestral angiosperm IRs (Zhu et al., 2016) as the references, we observed that IRs of most Convolvulaceous genera lack *rpl2* and *rpl23* (Figure 3). As a result, the most parsimonious scenario is that the common ancestor of Convolvulaceae experienced a contraction so that *rpl2* and *rpl23* were excluded from its IRs, followed by multiple rounds of lineage-specific expansion and contraction that shaped the IR diversity. For example, (1) *Erycibe* and Ipomoeeae *s.l.* have undergone a parallel expansion to include *ycf1*, *rps15*, *ndhH*, and partial *ndhA* in their IRs. This expansion

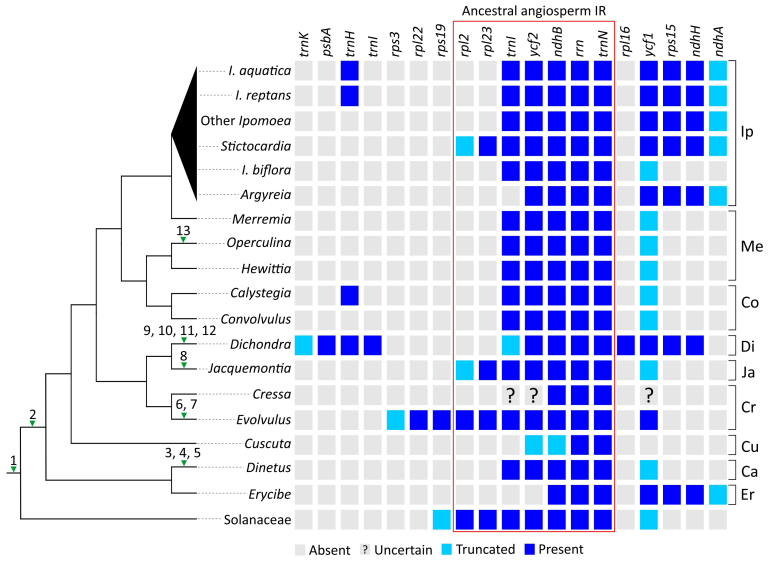


FIGURE 3
Evolution of plastomic rearrangements across Convolvulaceae. The right panel depicts presences/absences of genes in IRs. A "?" mark indicates that the gene state is uncertain due to poor sequence quality. Numerals on branches of the left panel tree denote specific rearrangements outside IRs. 1, loss of *infA*; 2, loss of the *rpl2* intron; 3, inversion of the *atpI–atpB* region; 4, duplication of *psbT*; 5, loss of *rpl23*; 6, loss of the *rps16* intron; 7, loss of the *rpoC1* intron; 8, integration of PTMTs between *psaA* and *ycf3*; 9, inversion of the *trnL–atpB* region; 10, loss of *rpl23*; 11, loss of the *rps16* intron; 12, loss of all introns from *ycf3*; 13, loss of *rpl23*. Ip, Ipomoeeae *s.l.*; Me, Merremieae; Co, Convolvuleae; Di, Dichondreae *s.l.*; Ja, Jacquemontieae; Cr, Cresseae *s.l.*; Cu, Cuscuteae; Ca, Cardiochlamyeae; Er, Erycibeae.

is here characterized as a synapomorphic trait of Ipomoeeae s.l., despite a subsequent contraction removing *rps15*, *ndhH*, and partial *ndhA* from *I. biflora*'s IRs; (2) *Dichondra* has duplicated *psbA*, *trnH*, *trnI*, *rpl16*, *ycf1*, *rps15*, *ndhH*, and partial *trnK* in its IRs, manifesting a complex IR expansion history including genes originally located in the LSC and SSC regions; (3) IR expansions have caused duplications of *rpl23*, *rpl2*, *rps19*, *rpl22*, and partial *rps3* in *Evolvulus* as well as *rpl23* and partial *rpl2* in *Jacquemontia* and *Sticotocardia*, whereas contractions have removed *trnI* and *ycf2* from the *Erycibe* IRs as well as *trnI* from the *Argyreia* IRs; (4) Small IR expansions have led to parallel duplications of *trnH* in *Calystegia*, *I. aquatica*, and *I. reptans*.

Integration of mtDNA in the *Jacquemontia* plastome

We detected an unusual elongated intergenic spacer between *psaA* and *ycf3* in the *Jacquemontia* plastome. This spacer is 9.1 kb long and approximately nine times longer than those (<1 kb) of other Convolvulaceous taxa. Furthermore, GC content of this intergenic spacer is elevated with cytosines and guanines accounting for 43.1% of all bases, in contrast to an average of 38.6% elsewhere (Figure 4A). Using this intergenic spacer sequence of *Jacquemontia* as the query, blast searches (NCBI nr database; access date: Jun/2022) yielded a mixture of fragments best matching mitogenomes of several taxa, including a basal angiosperm (*Amborella*) and other Convolvulaceous genera (Figure 4B). Such blast results, however, were not observed in other Convolvulaceous taxa whose syntenic sequences of this intergenic spacer always best matched plastomic sequences. Thus, this intergenic spacer between the *psaA* and *ycf3* genes of *Jacquemontia* could have been mis-assembled with contaminants from endogenous mtDNA, or may be a true integrated PTMT.

To verify the accuracy of our assembly, DNA-seq reads were mapped to the *Jacquemontia* plastome. Given the remarkably different copy numbers of plastomes compared to mitogenomes in plant cells (Johnston, 2019), we would expect drastic decreases of the read coverage in mis-assembled regions. However, our mapping analysis yielded a similar degree of the read coverage across the entire plastome (Figure 4A), thus rejecting the assumption that our assembly had endogenous mtDNA contamination. Both NOVOPlasty (Dierckxsens et al., 2017) and GetOrganelle (Jin et al., 2020) assemblers, though based on different assembly strategies, obtained identical results that again confirmed the presence of a PTMT in *Jacquemontia*. We also designed specific primers to amplify the region across the intergenic spacer between *psaA* and *ycf3*. Our PCR successfully yielded fragments of the expected intergenic size (9.1 Kb; Figure 4C). Collectively, our data demonstrate an unprecedented case of PTMTs in Convolvulaceae.

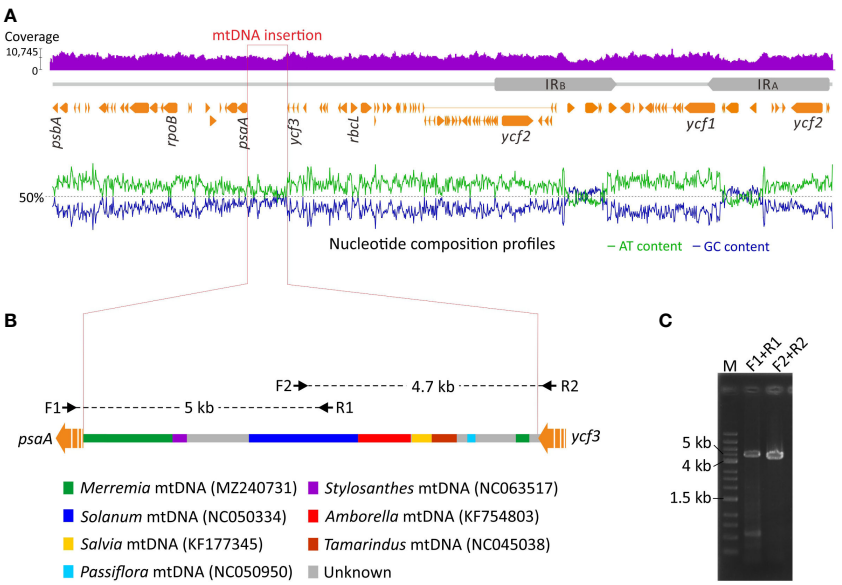


FIGURE 4
Schematic diagrams of PTMT integration in *Jacquemontia*. **(A)** Analyses of the DNA-seq read coverage and nucleotide composition across the whole plastome. **(B)** Blast results of the integrated PTMT showing a mixture of fragments best matching mtDNAs of some taxa with color codes given below. Black arrows are PCR primers designed in this study. **(C)** PCR verification of the integrated PTMT.

10.3389/fpls.2022.1061174

Short, domain-lacking, but repeat-containing introns in duplicated *rpl16* genes

As mentioned above, *Dichondra* has duplicated *rpl16* in its IRs, resulting in three *rpl16* copies in our assembly. One of these copies is located at the typical locus between *rpl14* and *rps3* in the LSC (hereafter designated as *rpl16*-LSC), while the others are located between *trnN* and *ycf1* in the IRs (hereafter designated as *rpl16*-IR). The two types of the *rpl16* copies differ in intron composition but share nearly identical coding sequences (% identical sites between different copies = 99.1). The intron of *rpl16*-LSC is 966 bp long with I–VI domains forming a typical group II intron structure (Figure 5A). In contrast, the *rpl16*-IR intron is relatively short (429 bp long) and completely lacks domains II–VI and EBS1 and EBS2 motifs in domain I. It is noteworthy that we detected a pair of direct repeats — one copy (205 bp long) in the region encompassing the exon 1 and its upstream region and the other (201 bp) at the 3'-end of the intron (Figure 5A). Their pairwise sequence identity is 97.6%.

The above-mentioned short, domain-lacking, but repeat-containing intron prompted us to ask two questions: (1) Is the intron in *rpl16*-IR still actively splicing? Specifically, we want to know if formation of an open reading frame is achieved after intron splicing. (2) Does homologous recombination between repeats occur and result in an intron-lacking *rpl16*-IR copy in

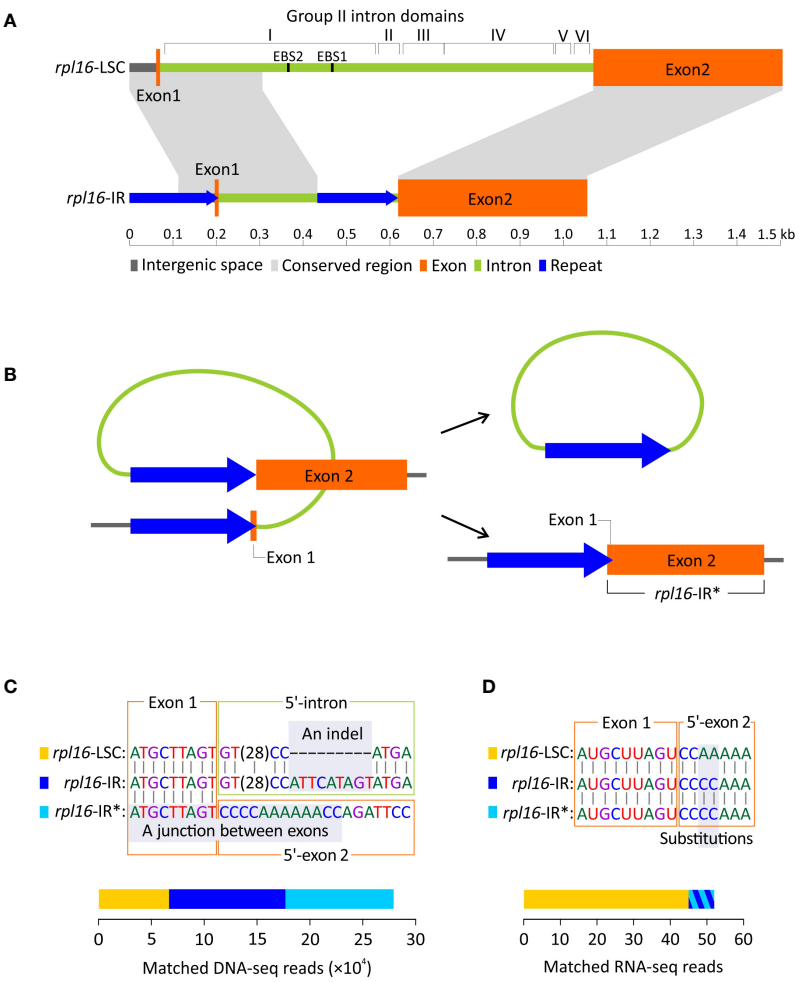


FIGURE 5
Duplication of *rpl16* in *Dichondra*. **(A)** Comparisons of *rpl16* copies residing in the LSC (*rpl16*-LSC) and IR (*rpl16*-IR) regions. The *rpl16*-LSC harbors a typical group II intron containing I–VI domains, while the *rpl16*-IR is short and degraded with only partial sequences of the domain I **(B)** A hypothetical scenario for loss of the *rpl16*-IR intron *via* homologous recombination between direct repeats. The intron-lacking *rpl16*-IR is therefore designated as *rpl16*-IR*. **(C)** DNA-seq mapping analyses showing the coexistence of three *rpl16* copies (color-coded) in *Dichondra*. Regions enabling copy discrimination are highlighted with grey and indicated with the terms "An indel" and "A junction between exons". Stacked bars are the number of reads matching specific copies. **(D)** RNA-seq reads mapped to putative transcripts of the three *rpl16* copies after splicing. The bars for *rpl16*-IR and *rpl16*-IR* are added together because they are not distinguishable by highlighted substitutions.

Dichondra (Figure 5B)? The intron-lacking copy (hereafter designated as *rpl16*-IR*) is separated from other two genomic *rpl16* copies by the junction of two exons (Figure 5C). Furthermore, the *rpl16*-LSC and *rpl16*-IR copies are distinguishable from each other by a 9 bp long indel at the 5'-end of their introns (Figure 5C). Our mapping analyses reveal 6,665, 10,228, and 11,032 DNA-seq reads that match *rpl16*-LSC, *rpl16*-IR, and *rpl16*-IR*, respectively (Figure 5C; Figure S3). These results confirm that (1) at least three different *rpl16* copies coexist in the *Dichondra* plastomes, (2) the genomic copy numbers of *rpl16*-IR and *rpl16*-IR* are approximately equal but two-fold higher than *rpl16*-LSC, consistent with the fact that the former two are in the IR, while the latter is single in the LSC, and (3) the equal numbers of matched reads between *rpl16*-IR and *rpl16*-IR* copies suggest effective repeat-mediated recombination. The coexistence and copy number divergence were further verified by semi-quantitative PCR that shows specific duplicate amplicons appearing earlier than the single-copy locus (Figure S3).

To examine splicing capability, RNA-seq reads were mapped to putative mature transcripts of the three *rpl16* copies (Figure 5D). We then detected 45 RNA reads that not only spanned exon junctions but also contained substitutions specific to *rpl16*-LSC. Eight RNA reads were found to match both *rpl16*-IR and *rpl16*-IR* because the mature transcripts of these two copies are identical, leading to difficulty in identification of which copies contributed to the matched reads. Therefore, it remains uncertain whether the *rpl16*-IR's intron is spliced. If it is, the splicing efficiency should be much lower than that in *rpl16*-LSC.

Discussion

To date, the genus *Ipomoea* is significantly overrepresented among sequenced morning glory plastomes. This bias holds back advances in understanding the evolution and utility of plastomes across the family. Our study compared plastomes representing nine of the 12 tribes, and provides new insights into phylogeny, plastome structural variation, PTMTs, and mechanisms underlying intron loss in the family.

Monophyletic status of the family, as well as phylogenetic placements of several of its tribes and genera, are still subject to debate. We adopted the classification system modified by Stefanovi´c et al. (2003), who recognized twelve tribes within Convolvulaceae rather than nine in Austin's system (1998). Notably, Austin (1998) raised the parasitic genus *Cuscuta* to the status of a monogeneric family (Cuscutaceae), but Stefanovi´c et al. (2003) considered it a monogeneric tribe (Cuscuteae). The present phylogenetic analyses reveal that *Cuscuta* is nested within Convolvulaceae, in line with the viewpoint that Convolvulaceae, including *Cuscuta*, is monophyletic (Stefanović et al., 2002; Stefanović and Olmstead, 2004; Lin et al., 2022; Simões et al., 2022). This monophyly is further reinforced by a common absence of the *rpl2* intron unique to Convolvulaceae among the subclass Asterideae (Stefanović et al., 2002; Lin et al., 2022; this study). However, our trees placed *Cuscuta* as a sister to other Convolvulaceous taxa except for the two genera *Dinetus* and *Erycibe* (Figure 1). This placement has never been reported before, even in the most recent plastid phylogenomic analysis by Lin et al. (2022). As our sampled taxa are denser and broader than those (25 taxa across eight tribes) in Lin et al. (2022), we conclude that the phylogenetic position of *Cuscuta* is unsettled and that expanded taxonomic sampling will be required to resolve *Cuscuta*'s evolutionary status.

Our analyses yielded strong support for a close relationship between Ipomoeeae *s.l.* and *Merremia* (a genus in the tribe Merremieae), confirming non-monophyly of the tribe Merremieae and calling for its revision. Furthermore, both of our ML (BS = 100%) and BI (PP = 1) trees resolve the clade comprising four genera: *Dichondra*, *Jacquemontia*, *Cressa*, and *Evolvulus*. Taxa in this clade, except *Jacquemontia*, possess deeply divided styles, the so-called "bifid style" clade or the subfamily Dicranostyloideae (Stefanović et al., 2002). Recently, the placement of *Jacquemontia* within Dicranostyloideae was recovered with full support in coalescent trees based on multiple nuclear loci (Simões et al., 2022). As a result, the placement of *Jacquemontia* within Dicranostyloideae is confirmed by a variety of methods and molecular data sets. However, the position of the genus *Erycibe* appears to be discordant among plastid, mitochondrial, and nuclear trees. For example, a sisterhood between *Dinetus* and *Erycibe* is weakly supported (BS=69%; PP = 0.994) in our trees (Figure 1), but such relationship was not observed in the trees inferred from mitochondrial CDSs and nuclear 45S (Lin et al., 2022). Simões et al. (2022) noted shifts of the *Erycibe* position with and without incorporation of *Cuscuta* in their analytical dataset, implicating the influence of taxonomic sampling. Collectively, our results affirm some but not all relationships in Convolvulaceae. The unresolved (or weakly supported) relationships are likely due to highly variable rates in nucleotide substitutions among taxa as demonstrated in Figure 2.

We identified numerous plastomic rearrangements in the LSC and IR regions (Figure 3). However, more than half (7/13 = 0.54) of the LSC rearrangements occurred in the subfamily Dicranostyloideae, indicating an asymmetric distribution of plastomic rearrangements across Convolvulaceae. Taxa in Dicranostyloideae also exhibit variable IRs and accelerated nucleotide substitution rates (Figure 2), implying an association between plastomic rearrangements and nucleotide substitution rates. This association likely results from aberrant mutation rates and/or improper DNA repair systems proposed for some Geraniaceae lineages whose plastomic rearrangements and nucleotide substitution rates are co-elevated (Guisinger et al., 2011; Blazier et al., 2016). IRs might act to stabilize plastomes (Palmer and Thompson, 1982). This hypothetical

function should be ineffective in Convolvulaceae. Blazier et al. (2016) suggested that plastome stabilization by IRs is evident when repeat content is low. Indeed, we have detected a pair of direct repeats and shown its capability of efficiently triggering recombination (Figure 5). We propose that multiple independent rounds of IR contraction and expansion have shaped the evolution of gene content in the IRs of Convolvulaceae. Illegitimate recombination between IR and single copy regions might have led to IR contraction and expansion (Goulding et al., 1996). Although not involved in gene losses, IR size fluctuation has resulted in copy-number variation of some genes in some Convolvulaceae plastomes (Figure 3). The low expression level we observed in *rpl16*-IR and *rpl16*-IR* implies down-regulation of these duplicates (Figure 5). Whether expression regulation also acts on other duplicate IR resident loci to prevent (or mediate) dosage effects remains to be elucidated.

In this study we report the first PTMT case in Convolvulaceae (Figure 4). This PTMT replaces the plastid intergenic sequences between *psaA* and *ycf3* in *Jacquemontia*. Replacements of plastid sequences with PTMTs were previously documented in *Daucus* (Iorizzo et al., 2012a; Iorizzo et al., 2012b), *Pariana* (Ma et al., 2015), and *Convallaria* (Raman et al., 2021). In *Jacquemontia*, the PTMT is inserted upstream of the *psaA* operon consisting of *psaA*, *psaB*, and *rps14* (Meng et al., 1988). This insertion has interrupted the transcription of the *psaA* operon driven by the original plastid promoters (Figure 5). Therefore, the expression machinery of the *psaA* operon must be reshaped after the PTMT insertion to ensure regular transcription of *psaA* and *psaB*, for their products are essential for the assembly of the PSI complex. In *Daucus*, replacements of plastid promoters with PTMTs were hypothesized to alter the expression pattern of the corresponding genes (Iorizzo et al., 2012a). A recent study suggested that mis-regulation of the plastid *psbB* operon results in hybrid incompatibility that may ultimately lead to speciation in *Oenothera* (Zupok et al., 2021). The PTMT we identified in *Jacquemontia* provides an opportunity to assess the association between PTMT insertions and speciation.

A variety of mechanisms can underlie PTMT insertions. The presence of short repeats and transcriptase-like sequences leads to the conclusion that the PTMT in *Daucus* was a consequence of transposition events governed by non-LTR retrotransposons (Iorizzo et al., 2012a; Iorizzo et al., 2012b). In *Asclepias*, migrations of mtDNA to plastids were achieved by homologous recombination between mitochondrial plastid-derived DNA and plastid sequences (Straub et al., 2013). Previously, deletions adjacent to insertion sites were considered to be signatures of homologous recombination facilitating PTMT integration in *Anacardium* (Rabah et al., 2017). Nevertheless, neither transposable elements nor deletions were detected in the PTMT or its flanking regions in *Jacquemontia*. Surprisingly, our blast searches revealed that the *Jacquemontia* PTMT contains several regions that best match mitogenomes of diverse species across five taxonomic orders (Amborellales, Fabales, Lamiales, Malpighiales, and Solanales; Figure 4). It is implausible to assume multiple rounds of horizontal gene transfer (HGT) from different origins into the same plastid intergenic region. Several studies have shown that PTMTs preserve the ancestral state of their mitochondrial donors (e.g., Iorizzo et al., 2012a; Iorizzo et al., 2012b; Straub et al., 2013; Ma et al., 2015). Unfortunately, we did not obtain any matches when the PTMT of *Jacquemontia* was used to blast against the mtDNA scaffolds we generated from conspecific PE reads. It is known that land plant mitogenomes can acquire abundant foreign DNA from intercellular gene transfer (IGT), HGT, or both (Knoop, 2004; Alverson et al., 2010; Rice et al., 2013; Park et al., 2014). Accordingly, we propose a straightforward scenario that the *Jacquemontia* PTMT has retained the ancestral state of its mitochondrial donor that once harbored an array of foreign DNA of different origins but later experienced reduction by purging the sequences homologous to this PTMT.

We found a short group II intron in the *rpl16*-IR copy (Figure 5). This intron is degraded and likely not functional since it lacks most of the domains, including the domain V for binding catalytic Mg^{2+} during intron splicing (Lambowitz and Zimmerly, 2011). This degraded intron is flanked by direct repeats capable of triggering homologous recombination to precisely remove it. This repeat-mediated recombination happens with a rate of 0.52 [11,032/(10,228 + 11,032)] conversions per copy and its associated product is transcribable, albeit at a low level. These findings lead us to speculate that these direct repeats might have been selected to complement the degraded intron since the intron is nonfunctional.

Retroprocessing is a well-known mechanism driving intron loss in plant organelles through conversion between retro-transcribed cDNA molecules and their corresponding DNA fragments. This mechanism is especially prominent when transcripts of target genes lack RNA editing in regions adjacent to former splicing sites (Ran et al., 2010; Sloan et al., 2010; Grewe et al., 2011; Cuenca et al., 2016). In addition, HGT and sequential gene conversion have created a chimeric *cox2* gene without introns (Hepburn et al., 2012). Here, we provide a different model for intron loss after invasion of direct repeats at appropriate positions in intron-containing genes. This model better interprets intron loss from non-chimeric genes that have abundant RNA editing sites in regions adjacent to the former intron splicing sites.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: NCBI, LC729542-LC729560 and the SRA bioproject: PRJNA876072.

Author contributions

S-MC and C-SW conceived and designed the study. C-SW performed experiments and data analyses. C-SW and C-IC collected the plant materials. C-SW and S-MC wrote the manuscript. All the authors checked and approved the final version. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by research grants from National Science and Technology Council, Taiwan (108-2621-B-001-006-) and Biodiversity Research Center, Academia Sinica (AS), Taiwan, and partially from AS-23-23 of AS president office to S-MC.

Acknowledgments

We thank Taipei Botanical Garden for providing some plant materials. We also thank Wayne Y. Lin and two reviewers for their critical reading and constructive comments.

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.2022.1061174/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Plastid phylogenomics of the tribe Ipomoeeae *s.l.* The tree details the simplified area depicted in . A 50% majority rule was used to condense the tree topology. Values along branches are bootstrap supports (%) and posterior possibilities for ML and Bayesian inference (BI) trees, respectively. Conflicting topologies are highlighted with grey.

SUPPLEMENTARY FIGURE 2
Dating of taxon divergence time during Convolvulaceae evolution. Black circles indicate nodes with constrained ages. Blue bars denote a range of the 95% highest posterior density (HPD).

SUPPLEMENTARY FIGURE 3
Read-mapping and Semi-quantitative PCR demonstrating coexistence of the three different *rpl16* copies and discrepancies in their copy numbers in *Dichondra*. Primers, *rpl16*-LSC-F and *rpl16*-LSC-R, were designed to amplify the region containing intronic II–IV domains that are lacking in the *rpl16*-IR copy.

References

Alverson, A. J., Wei, X., Rice, D. W., Stern, D. B., Barry, K., and Palmer, J. D. (2010). Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). *Mol. Biol. Evol.* 27, 1436–1448. doi: 10.1093/molbev/msq029

Amiryousefi, A., Hyvönen, J., and Poczai, P. (2018). The chloroplast genome sequence of bittersweet (*Solanum dulcamara*): Plastid genome structure evolution in solanaceae. *PloS One* 13, e0196069. doi: 10.1371/journal.pone.0196069

Angiosperm Phylogeny Group (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

Austin, D. F. (1998). "Parallel and convergent evolution in the convolvulaceae," in *Biodiversity and taxonomy of tropical flowering plants*. Eds. P. Mathews and M. Sivadasan (Calicut: Mentor Books), 201–234.

Banerjee, A., and Stefanović, S. (2020). Reconstructing plastome evolution across the phylogenetic backbone of the parasitic plant genus *Cuscuta* (Convolvulaceae). *Biol. J. Linn. Soc Lond.* 194, 423–438. doi: 10.1093/botlinnean/boaa056

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

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

Braukmann, T., Kuzmina, M., and Stefanović, S. (2013). Plastid genome evolution across the genus *Cuscuta* (Convolvulaceae): two clades within subgenus *Grammica* exhibit extensive gene loss. *J. Exp. Bot.* 64, 977–989. doi: 10.1093/jxb/ers391

Chan, P. P., Lin, B. Y., Mak, A. J., and Lowe, T. M. (2021). tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* 49, 9077–9096. doi: 10.1093/nar/gkab688

Cuenca, A., Ross, T. G., Graham, S. W., Barrett, C. F., Davis, J. I., Seberg, O., et al. (2016). Localized retroprocessing as a model of intron loss in the plant mitochondrial genome. *Genome Biol. Evol.* 8, 2176–2189. doi: 10.1093/gbe/evw148

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

Downie, S. R., Olmstead, R. G., Zurawski, G., Soltis, D. E., Soltis, P. S., Watson, J. C., et al. (1991). Six independent losses of the chloroplast DNA rpl2 intron in dicotyledons: molecular and phylogenetic implications. *Evolution.* 45, 1245–1259. doi: 10.1111/j.1558-5646.1991.tb04390.x

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

Eserman, L. A., Tiley, G. P., Jarret, R. L., Leebens-Mack, J. H., and Miller, R. E. (2014). Phylogenetics and diversification of morning glories (tribe ipomoeeae, convolvulaceae) based on whole plastome sequences. *Am. J. Bot.* 101, 92–103. doi: 10.3732/ajb.1300207

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

Grewe, F., Herres, S., Viehöver, P., Polsakiewicz, M., Weisshaar, B., and Knoop, V. (2011). A unique transcriptome: 1782 positions of RNA editing alter 1406 codon identities in mitochondrial mRNAs of the lycophyte *Isoetes engelmannii*. *Nucleic Acids Res.* 39, 2890–2902. doi: 10.1093/nar/gkq1227

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. *Mol. Biol. Evol.* 28, 583–600. doi: 10.1093/molbev/msq229

Hepburn, N. J., Schmidt, D. W., and Mower, J. P. (2012). Loss of two introns from the *Magnolia tripetala* mitochondrial *cox2* gene implicates horizontal gene transfer and gene conversion as a novel mechanism of intron loss. *Mol. Biol. Evol.* 29, 3111–3120. doi: 10.1093/molbev/mss130

Iorizzo, M., Grzebelus, D., Senalik, D., Szklarczyk, M., Spooner, D., and Simon, P. (2012a). Against the traffic: The first evidence for mitochondrial DNA transfer into the plastid genome. *Mob. Genet. Elements* 2, 261–266. doi: 10.4161/mge.23088

Iorizzo, M., Senalik, D., Szklarczyk, M., Grzebelus, D., Spooner, D., and Simon, P. (2012b). *De novo* assembly of the carrot mitochondrial genome using next generation sequencing of whole genomic DNA provides first evidence of DNA transfer into an angiosperm plastid genome. *BMC Plant Biol.* 12, 61. doi: 10.1186/1471-2229-12-61

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

Johnston, I. G. (2019). Tension and resolution: dynamic, evolving populations of organelle genomes within plant cells. *Mol. Plant* 12, 764–783. doi: 10.1016/j.molp.2018.11.002

Kelchner, S. A. (2002). Group II introns as phylogenetic tools: structure, function, and evolutionary constraints. *Am. J. Bot.* 89, 1651–1669. doi: 10.3732/ajb.89.10.1651

Knoop, V. (2004). The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. *Curr. Genet.* 46, 123–139. doi: 10.1007/s00294-004-0522-8

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Lambowitz, A. M., and Zimmerly, S. (2011). Group II introns: mobile ribozymes that invade DNA. cold spring harb. *Perspect. Biol.* 3, a003616. doi: 10.1101/cshperspect.a003616

Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., and Calcott, B. (2017). PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol. Biol. Evol.* 34, 772–773. doi: 10.1093/molbev/msw260

Laux, M., Oliveira, R. R. M., Vasconcelos, S., Pires, E. S., Lima, T. G. L., Pastore, M., et al. (2022). New plastomes of eight ipomoea species and four putative hybrids from Eastern Amazon. *PloS One* 17, e0265449. doi: 10.1371/journal.pone.0265449

Lin, Y., Li, P., Zhang, Y., Akhter, D., Pan, R., Fu, Z., et al. (2022). Unprecedented organelle genomic variations in morning glories reveal independent evolutionary scenarios of parasitic plants and the diversification of plant mitochondrial complexes. *BMC Biol.* 20, 49. doi: 10.1186/s12915-022-01250-1

Ma, P. F., Zhang, Y. X., Guo, Z. H., and Li, D. Z. (2015). Evidence for horizontal transfer of mitochondrial DNA to the plastid genome in a bamboo genus. *Sci. Rep.* 5, 11608. doi: 10.1038/srep11608

Meng, B. Y., Tanaka, M., Wakasugi, T., Ohme, M., Shinozaki, K., and Sugiura, M. (1988). Cotranscription of the genes encoding two P700 chlorophyll a apoproteins with the gene for ribosomal protein CS14: determination of the transcriptional initiation site by *in vitro* capping. *Curr. Genet.* 14, 395–400. doi: 10.1007/BF00419998

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., et al. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37, 1530–1534. doi: 10.1093/molbev/msaa015

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., Ruhlman, T. A., Sabir, J. S., Mutwakil, M. H., Baeshen, M. N., Sabir, M. J., et al. (2014). Complete sequences of organelle genomes from the medicinal plant *Rhazya stricta* (Apocynaceae) and contrasting patterns of mitochondrial genome evolution across asterids. *BMC Genomics* 15, 405. doi: 10.1186/1471-2164-15-405

Park, I., Yang, S., Kim, W. J., Noh, P., Lee, H. O., and Moon, B. C. (2018). The complete chloroplast genomes of six ipomoea species and indel marker development for the discrimination of authentic pharbitidis semen (seeds of *I. nil* or *I. purpurea*). *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00965

Rabah, S. O., Lee, C., Hajrah, N. H., Makki, R. M., Alharby, H. F., and Alhebshi, A. M. (2017). Plastome sequencing of ten nonmodel crop species uncovers a large insertion of mitochondrial DNA in cashew. *Plant Genome.* 10. doi: 10.3835/plantgenome2017.03.0020

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., 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

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A.. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* 67, 901–904. doi: 10.1093/sysbio/syy032

Ran, J. H., Gao, H., and Wang, X. Q. (2010). Fast evolution of the retroprocessed mitochondrial rps3 gene in conifer II and further evidence for the phylogeny of gymnosperms. *Mol. Phylogenet. Evol.* 54, 136–149. doi: 10.1016/j.ympev.2009.09.011

Rice, D. W., Alverson, A. J., Richardson, A. O., Young, G. J., Sanchez-Puerta, M. V., Munzinger, J., et al. (2013). Horizontal transfer of entire genomes *via* mitochondrial fusion in the angiosperm *Amborella*. *Science.* 342, 1468–1473. doi: 10.1126/science.1246275

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., et al. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. doi: 10.1093/sysbio/sys029

Simões, A. R. G., Eserman, L. A., Zuntini, A. R., Chatrou, L. W., Utteridge, T. M. A., Maurin, O., et al. (2022). A bird's eye view of the systematics of convolvulaceae: novel insights from nuclear genomic data. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.889988

Sloan, D. B., MacQueen, A. H., Alverson, A. J., Palmer, J. D., and Taylor, D. R. (2010). Extensive loss of RNA editing sites in rapidly evolving *Silene* mitochondrial genomes: selection vs. retroprocessing as the driving force. *Genetics* 185, 1369–80. doi: 10.1534/genetics.110.118000

Smith, D. R. (2011). Extending the limited transfer window hypothesis to inter-organelle DNA migration. *Genome Biol. Evol.* 3, 743–748. doi: 10.1093/gbe/evr068

Stefanović, S., Austin, D. F., and Olmstead, R. G. (2003). Classification of convolvulaceae: a phylogenetic approach. *Syst. Bot.* 28, 791–806. doi: 10.1043/02-45.1

Stefanović, S., Krueger, L., and Olmstead, R. G. (2002). Monophyly of the convolvulaceae and circumscription of their major lineages based on DNA sequences of multiple chloroplast loci. *Am. J. Bot.* 89, 1510–1522. doi: 10.3732/ajb.89.9.1510

Stefanović, S., and Olmstead, R. G. (2004). Testing the phylogenetic position of a parasitic plant (*Cuscuta*, convolvulaceae, asteridae): Bayesian inference and the parametric bootstrap on data drawn from three genomes. *Syst. Biol.* 53, 384–399. doi: 10.1080/10635150490445896

Stewart, C. N.Jr., and Via, L. E. (1993). A rapid CTAB DNA isolation technique useful for RAPD fingerprinting and other PCR applications. *Biotechniques* 14, 748–750.

Straub, S. C., Cronn, R. C., Edwards, C., Fishbein, M., and Liston, A. (2013). Horizontal transfer of DNA from the mitochondrial to the plastid genome and its subsequent evolution in milkweeds (Apocynaceae). *Genome Biol. Evol.* 5, 1872–1885. doi: 10.1093/gbe/evt140

Sun, J., Dong, X., Cao, Q., Xu, T., Zhu, M., Sun, J., et al. (2019). A systematic comparison of eight new plastome sequences from ipomoea l. *Peer J.* 7, e6563. doi: 10.7717/peerj.6563

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One* 9, e112963. doi: 10.1371/journal.pone.0112963

Wang, R. J., Cheng, C. L., Chang, C. C., Wu, C. L., Su, T. M., and Chaw, S. M. (2008). Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. *BMC Evol. Biol.* 8, 36. doi: 10.1186/1471-2148-8-36

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

Wu, C. S., Sudianto, E., and Chaw, S. M. (2021). Tight association of genome rearrangements with gene expression in conifer plastomes. *BMC Plant Biol.* 21, 33. doi: 10.1186/s12870-020-02809-2

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

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

Zupok, A., Kozul, D., Schöttler, M. A., Niehörster, J., Garbsch, F., Liere, K., et al. (2021). A photosynthesis operon in the chloroplast genome drives speciation in evening primroses. *Plant Cell.* 33, 2583–2601. doi: 10.1093/plcell/koab155

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 03 March 2023
DOI 10.3389/fpls.2023.1070600

OPEN ACCESS

EDITED BY
Peter Poczai,
University of Helsinki, Finland

REVIEWED BY
Emre Sevindik,
Adnan Menderes University, Türkiye
Woojong Jang,
Herbal Medicine Research Department,
Korea Institute of Oriental Medicine
(KIOM), Republic of Korea

*CORRESPONDENCE
Yu-liang Cai
yuanyicyl@nwafu.edu.cn

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

RECEIVED 15 October 2022
ACCEPTED 21 February 2023
PUBLISHED 03 March 2023

CITATION
Wan T, Qiao B-x, Zhou J, Shao K-s,
Pan L-y, An F, He X-s, Liu T, Li P-k and
Cai Y-l (2023) Evolutionary and
phylogenetic analyses of 11 *Cerasus*
species based on the complete
chloroplast genome.
Front. Plant Sci. 14:1070600.
doi: 10.3389/fpls.2023.1070600

Evolutionary and phylogenetic analyses of 11 *Cerasus* species based on the complete chloroplast genome

Tian Wan[1], Bai-xue Qiao[1], Jing Zhou[1], Ke-sen Shao[1], Liu-yi Pan[1], Feng An[1], Xu-sheng He[2], Tao Liu[1], Ping-ke Li[3] and Yu-liang Cai[1]*

[1]College of Horticulture, Northwest Agriculture & Forestry University, Yangling, China, [2]College of Natural Resources and Environment, Northwest Agriculture & Forestry University, Yangling, China, [3]Center of Experimental Station, Northwest Agriculture & Forestry University, Yangling, China

The subgenus *Cerasus*, one of the most important groups in the genus *Prunus sensu lato*, comprises over 100 species; however, the taxonomic classification and phylogenetic relationships of *Cerasus* remain controversial. Therefore, it is necessary to reconstruct the phylogenetic tree for known *Cerasus* species. Here, we report the chloroplast (cp) genome sequences of 11 *Cerasus* species to provide insight into evolution of the plastome. The cp genomes of the 11 *Cerasus* species (157,571–158,830 bp) displayed a typical quadripartite circular structure. The plastomes contain 115 unique genes, including 80 protein-coding genes, four ribosomal RNAs, and 31 transfer RNAs. Twenty genes were found to be duplicated in inverted repeats as well as at the boundary. The conserved non-coding sequences showed significant divergence compared with the coding regions. We found 12 genes and 14 intergenic regions with higher nucleotide diversity and more polymorphic sites, including *matK*, *rps16*, *rbcL*, *rps16-trnQ*, *petN-psbM*, and *trnL-trnF*. During cp plastome evolution, the codon profile has been strongly biased toward the use of A/T at the third base, and leucine and isoleucine codons appear the most frequently. We identified strong purifying selection on the *rpoA*, *cemA*, *atpA*, and *petB* genes; whereas *ccsA*, *rps19*, *matK*, *rpoC2*, *ycf2* and *ndhI* showed a signature of possible positive selection during the course of *Cerasus* evolution. In addition, we further analyzed the phylogenetic relationships of these species with 57 other congenic related species.Through reconstructing the *Cerasus* phylogeny tree, we found that true cherry is similar to the flora of China forming a distinct group, from which *P. mahaleb* was separated as an independent subclade. *Microcerasus* was genetically closer to *Amygdalus*, *Armeniaca*, and *Prunus* (*sensu stricto*) than to members of true cherry, whereas *P. japonica* and *P. tomentosa* were most closely related to *P. triloba* and *P. pedunculata*. However, *P. tianshanica* formed a clade with *P. cerasus*, *P. fruticosa*, *P. cerasus* × *P. canescens* 'Gisela 6', and *P. avium* as a true cherry group. These results provide new insights into the plastome evolution of *Cerasus*, along with potential molecular markers and candidate DNA barcodes for further phylogenetic and phylogeographic analyses of *Cerasus* species.

KEYWORDS

***Prunus*, plastome, genomic variation, repeat sequence, protein-coding gene, phylogeny**

1 Introduction

Rosaceae is a large family that includes most economically important fruits species in temperate zones, such as *Prunus*, *Armeniaca*, *Amygdalus*, *Pyrus*, *Malus* and *Crataegus* species (Zarei et al., 2017; Sevindik et al., 2020). And the plant subgenus *Cerasus* is considered one of the most important groups in the genus *Prunus (P.) sensu lato*, comprising over 100 species, which are naturally distributed in temperate Asia, Europe, North America, China, Japan, and Korea (Chin et al., 2014; Zhang et al., 2021). In China, there are roughly 45 species of *Cerasus*, 35 of which are considered to be endemic according to the Flora of China project (Yü et al., 1986). However, the taxonomic classification and phylogenetic relationships among species in the subgenus *Cerasus* or genus *Prunus sensu lato* have been controversial, with no unification reached to date (Potter et al., 2007; Chin et al., 2014; Liang et al., 2018; Zhang et al., 2021). For example, *P. tomentosa*, *P. tianshanica*, *P. japonica*, *P. humilis*, *P. dictyoneura*, *P. glandulosa*, *P. pogonostyla*, *P. jacquemontii*, *P. prostrata*, and *P. pumila* were classified in a single group, the "dwarf cherry" (*Microcerasus*), which was identified as a section of subgenus *Cerasus* according to Yü et al. (1986) and Webster and Looney (1996). However, in previous phylogenetic studies, *Microcerasus* species corresponded with *Amygdalus*, *Armeniaca*, or *Prunus* species (Bortiri et al., 2001; Shaw and Small, 2004; Chen et al., 2018; Zhang et al., 2021). Therefore, it is necessary to reconstruct the phylogenetic tree for *Cerasus* species.

The origin of the chloroplast (cp) can be traced back more than one billion years (Wang et al., 2021; Ravi et al., 2007). In land plants, the cp genome has a relatively conserved quadripartite structure, with conserved sequences in the range of 120–218 kb (Daniell et al., 2016) encoding approximately 100–130 genes (Palmer, 1985; Ravi et al., 2007; Wicke et al., 2011; Daniell et al., 2016; Wang et al., 2021). The cp structure comprises one large single-copy (LSC) region, one small single-copy (SSC) region, and two copies of an inverted repeat (IR) (Sugiura, 1992; Ravi et al., 2007). The significant developmental impact and limited coding potential of the cp genome, combined with maternal inheritance of this organelle provide tangible, causal approaches to understanding plant evolution, diversity, and phylogenic relationships (Ravi et al., 2007; Moore et al., 2010; Greiner et al., 2011). Recent studies have demonstrated that cp genome sequences offer remarkable resolution for analyzing phylogenetic relationships at various taxonomic levels, and can further provide evidence to explain effects of geography and climate oscillations on genetic divergence (Ivanova et al., 2017; Xue et al., 2019; Xue et al., 2021; Zhang et al., 2021; Dunning et al., 2022; Wang et al., 2022). Although cp genome sequences of some *Cerasus* species have been published (Chen et al., 2018; Feng et al., 2018; Zhang et al., 2021; Li et al., 2022), there is still a lack of information to enable comprehensive analysis of the interspecific relationships of the subgenus *Cerasus* and the relationship between *Cerasus* and *Prunus sensu lato*. Comparison of the cp genomes of the 11 *Cerasus* species can help to better understand evolution of the *Cerasus* genome and enable more profound analysis of the phylogenetic relationships in the genus *Prunus*, offering valuable insights.

In this study, we performed a comparative analysis of 11 complete *Cerasus* cp genomes to explore the features and structural differentiation of sequences among species. Furthermore, we reconstructed a phylogenetic tree using the newly obtained cp chloroplast sequences and published sequences to explore the genetic relationships among subgenus *Cerasus*, *Prunus* (*sensu stricto*), *Amygdalus*, and *Armeniaca*. Our study objectives were to: (1) gain insight into plastome structure features, (2) inform an improved understanding of cp genome evolution, and (3) further delineate the taxonomic status of *Cerasus*.

2 Materials and methods

2.1 Plant materials, sequencing, cp genome assembly, and annotation

Fresh and healthy leaves were collected from adult plants of 11 *Cerasus* species (Table 1). All samples were immediately frozen in liquid nitrogen and stored at −80°C until analysis. Total genomic DNA was extracted from 100 mg of fresh leaves using a modified CTAB method (Murray and Thompson, 1980). DNA libraries were prepared and sequenced on an Illumina NovaSeq 6000 platform (Illumina, San Diego, CA, USA) with paired-end 150-bp sequencing reads; only reads with a Q30 quality score greater than 80% were retained for analysis.

The cp genome assembly of *Cerasus* species was obtained by a baiting and iterative mapping approach (Hahn et al., 2013). The complete cp genome of *P. persica* (HQ336405) was downloaded from the National Center for Biotechnology Information (NCBI) database as a reference genome. Geneious Prime v2022.0.2 (https://www.geneious.com/) (Kearse et al., 2012) was used for sequence correction. The 11 *Cerasus* species were annotated by Geneious Prime v2022.0.2, using *P. pseudocerasus* (NC030599) and *P. persica* (HQ336405) as reference sequences, and were annotated using GeSeq (https://chlorobox.mpimp-golm.mpg.de/geseq.html) with no reference sequence. Manual editing of annotated and non-annotated portions, including exons and introns, was then performed. The transfer RNA (tRNA) sequences were confirmed using tRNAscan-SE 2.0 (Chan et al., 2021). All annotations were checked against the reference genomes (NC030599, NC054254, and MZ145044). Genome maps were drawn using OrganellarGenomeDRAW (OGDRAW) (Greiner et al., 2019).

2.2 Complete cp genome comparison

Plastome structures among *Cerasus* species, apple, pear, and grape were compared by the mVISTA percent identity plot in Shuffle-LAGAN mode to reveal the major genomic variations located in LSC and SSC regions (Brudno et al., 2003; Frazer et al., 2004). Subsequently, nucleotide diversity (*Pi*) and polymorphic sites (*S*) of single-copy genes and intergenic regions (IGRs) were estimated for the 11 species and *P. pseudocerasus* (NC030599) by DnaSP v.6 (Rozas et al., 2017). The plastome genetic architecture of the 11 *Cerasus* species, 24 *Cerasus* species available in NCBI, and six

TABLE 1 Sampling information for the *Cerasus* species.

No.	Species	Origin	Sampling sites, longitude, latitude	GenBank number
1	*P. avium* (wild)	Hungary	Mei County, Shaanxi, China E 107.9908°, N 34.1123°	OP598110
2	*P. cerasus*	Hungary	Mei County, Shaanxi, China E 107.9908°, N 34.1123°	MW477432
3	*P. cerasus* × *P. canescens* 'Gisela 6'	Germany	Qishan County, Shaanxi, China E 107.6371°, N 34.3749°	MW477433
4	*P. fruticosa*	Hungary	Mei County, Shaanxi, China E 107.9908°, N 34.1123°	MW477434
5	*P. japonica*	Shanxi, China	Hongdong, Shanxi, China E 111.8234°, N 36.4298°	OP598111
6	*P. mahaleb*	Hungary	Mei County, Shaanxi, China E 107.9908°, N 34.1123°	MW477435
7	*P. serrula*	Yunnan, China	Mei County, Shaanxi, China E 107.9908°, N 34.1123°	MW477436
8	*P. serrulata*	Shandong, China	RiZhao, Shandong, China E 119.2087°, N 35.7501°	OP611546
9	*P. tianshanica*	Xinjiang, China	Yili, Xinjiang, China E 81.2771°, N 43.9094°	OP598112
10	*P. tomentosa*	Shaanxi, China	Taibai, Shaanxi, China E 107.5947°, N 34.0533°	MW477437
11	*P. trichostoma*	Xizang, China	Nyingchi, Xizang, China E 94.6609°, N 29.6340°	OP598113

other species for the LSC/IR and SSC/IR boundaries were analyzed by Irscope (Amiryousefi et al., 2018).

2.3 Repeat sequences analysis

Simple sequence repeats (SSRs) were examined by the Perl script MicroSAtellite (MISA) (Beier et al., 2017) with the following parameter settings: motif size of 1–6 nucleotides; and minimum repeat unit of 10 for mononucleotides, 6 for dinucleotides, and 5 for tri-, tetra-, penta-, and hexa-nucleotides. Non-overlapping repeat sequences were identified by REPuter (Kurtz et al., 2001) (repeat unit length minimum ≥ 25 bp, Hamming distance = 3). Four matches of repeats were classified, namely forward, reverse, complement, and palindromic matches. The online program Tandem Repeats Finder (http://tandem.bu.edu/trf/trf.html) was used to find the tandem repeat sequences of at least 10 bp in length. The alignment parameters for match, mismatch, and indels were set to 2, 7, and 7, respectively.

2.4 Codon usage bias and gene selective pressure analysis

For identification of codon usage patterns, all coding sequences (CDSs) greater than 350 nucleotides in length were extracted from the cp genome of *Cerasus*, as described previously (Morton, 1998). The filtered CDSs were subsequently used for the estimation of codon usage using CodonW v.9.1.2 and the codon usage patterns were analyzed by GraphPad Prism v.8.0. In addition to the overall codon usage, we further tabulated codon usage measures such as the effective number of codons (Nc) and GC frequency at the third synonymous position (GC3s). To investigate the selective pressure on plastome protein-coding genes between two species, non-synonymous (Ka) and synonymous (Ks) substitution values were calculated by KaKs_Calculator 2.0 (Wang et al., 2010), with the following settings: genetic code table = 11 (bacterial and plant plastid codes) and the Yang-Nielsen algorithm (YN) calculation method (Ivanova et al., 2017).

2.5 Phylogenetic relationship

To reconstruct the phylogenetic relationships and verify the phylogenetic position of the subgenus *Cerasus* in *Prunus sensu lato*, 45 *Cerasus* cp genome sequences, including 34 published sequences downloaded from NCBI and the 11 sequences obtained in this study, were analyzed along with 23 *P.sensu lato* complete cp genome sequences. *Malus baccata*, *Malus micromalus*, *Pyrus communis*, *Pyrus ussuriensis*, *Vitis amurensis*, *V. amurensis*, and *Ziziphus jujuba* were regarded as outgroups. Because the different regions of cp genome differed in the molecular evolutionary rates (Zhang et al., 2017), phylogenetic relationship analyses were performed using 6 datasets that are the complete cp genome sequences, LSC regions, SSC regions, two IR regions, common CDS and IGRs. Sequences were aligned by MAFFT using Geneious Prime v2022.0.2

(Kearse et al., 2012). The phylogenetic tree was constructed with the program Mrbayes (Huelsenbeck and Ronquist, 2001) of Geneious Prime and Maximum likelihood (ML) method of MEGA v11 (Kumar et al., 2016), Mrbayes analysis used the following Markov chain Monte Carlo simulation settings: chain length = 1,100,000, subsampling frequency = 200, heated chains = 4, burn-in length = 100,000, heated chain temp = 0.2, and random seed = 170. ML analyses used the General Time Reversible +Gamma Distributed + Nearest-Neighbor-Interchange model with 1000 bootstrap replicates. The tree was visualized with Interactive Tree Of Life (iTOL) v5 software (Letunic and Bork, 2021) and manually edited where necessary.

3 Results

3.1 Cp genome structure of *Cerasus* species

The *Cerasus* cp genome displays a typical quadripartite circular structure (Figure 1) containing one LSC, one SSC, and two IR (IRB and IRA) regions, as determined by OGDRAW (Greiner et al., 2019). The plastome size of the 11 *Cerasus* species ranged from 157,571 bp (*P. tomentosa*) to 158,830 bp (*P. tianshanica*). The average coverage depth ranged from 421.8× to 9410× (Table S1; Figure S1). The GC content of the *Cerasus* cp genome was very similar among the 11 species (36.5–36.8%) with an average of 36.7% (Table S1).

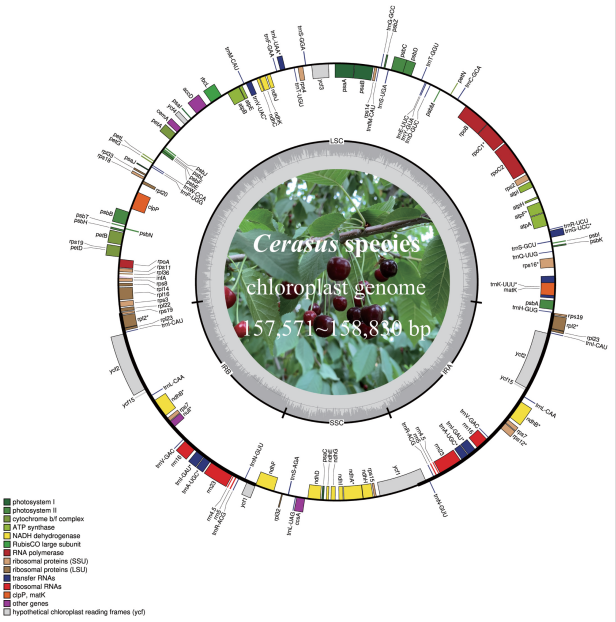


FIGURE 1
Chloroplast genome map of *Cerasus* species. Genes on the inside of the large circle are transcribed clockwise and those on the outside are transcribed counterclockwise. The genes are color-coded based on their functions. The dashed area represents the GC composition of the chloroplast genome. LSC, large single copy region; IR, inverted repeat; SSC, small single copy region.

Cerasus plastomes contained the same set of 115 unique genes, including 80 protein-coding genes, 4 ribosomal RNAs (rRNAs), and 31 tRNAs (Table 2). Twenty genes are duplicated in IRs or at the boundaries, including nine protein-coding genes (*rpl2*, *rpl23*, *rps7*, *rps12*, *rps19*, *ndhB*, *ycf1*, *ycf2*, and *ycf15*), four rRNA genes (*rrn4.5*, *rrn5*, *rrn16*, and *rrn23*), and seven tRNA genes (*trnA-UGC*, *trnI-CAU*, *trnI-GAU*, *trnL-CAA*, *trnN-GUU*, *trnR-ACG*, and *trnV-GAC*). In the cp genome of *P. tianshanica*, an insertion sequence was identified that split *ycf2* in IRA into two segments (Figure S2A). There are 18 different intron-containing genes (Table 2), including 10 protein-coding genes (*atpF*, *ndhA*, *ndhB*, *rpl2*, *rpl16*, *rps12*, *rps16*, *rpoC1*, *petB*, and *petD*) and six tRNA-coding genes (*trnA-UGC*, *trnG-UCC*, *trnI-GAU*, *trnK-UUU*, *trnL-UAA*, and *trnV-UAC*). Among these, *trnK-UUU* has the largest intron (2428–2539 bp) with *matK* located within it. Two protein-coding genes (*clpP* and *ycf3*) contain two introns.

In *Cerasus* plastomes, the protein-coding gene *rps19* is located on the boundary of the LSC and IR regions, except for *P. avium* in which only the *rps19* fragment is on IRA. Comparison showed significant differences of *rps19* gene sequences among the *Cerasus* species (Figure S2B). The *ycf1* gene was located on the boundary of the SSC and IRA regions. *ycf* genes were identified as hypothetical cp reading frames, and small fragments of truncated *ycf* genes were detected in IRA (*ycf15*: 126 bp and 129 bp, respectively), with only partial *ycf1* identified in the IRB region. The remaining *ycf* genes were detected at the complete gene size. *ycf2* is a large functional gene encoding 2277 amino acids in cp IR regions. The *ycf2* gene in *P. tianshanica* was 6876 bp, with one inserted fragment of 42 bp located at the 900-bp position (Figure S2A). In *Cerasus*, a high level of similarity was restricted to the IRs, and major differences originated from the LSC and SSC regions. The gene *infA*, which is a translation-related gene, was identified as a pseudogene.

3.2 Complete cp genome sequence comparison of 11 *Cerasus* species

The mVISTA (Frazer et al., 2004) analysis showed the overall sequence identity, divergent regions, and visualization of the aligned cp genome sequences in *Cerasus*. The LSC and SSC regions were clearly more divergent than the IRs (Figure 2). The conserved non-coding sequences (CNSs) showed significantly more divergence than the coding regions (Figure 2 and Table S2), indicating that the CDSs are much more conserved than the CNSs. Furthermore, the mean value of *Pi* in IRs (0.00157) was lower than that of the LSC (0.00719) or SSC (0.01041) regions, which demonstrated that the IR regions have fewer mutations and are thus more strongly conserved. Among 135 plastid genes, only 23 genes showed higher nucleotide diversity ($Pi > 0.003$), with *Pi* values ranging from 0.00307 (*rpoC2*) to 0.00777 (*rps15*), and 12 genes (*trnK-UUU*, *matK*, *rps16*, *trnG-UCC*, *rpoC2*, *rbcL*, *accD*, *clpP*, *rpl16*, *ndhF*, *ndhA*, and *ycf1*) had a relatively higher number of polymorphic sites ($S > 10$) (Table S2A; Figure S3). However, 73 IGRs had $Pi > 0.003$; the 10 most polymorphic IGRs in ascending order were *rps19-trnG-GUU* ($Pi = 0.06926$), *trnR-UCU-atpA*, *ndhC-trnV-UAC*, *ccsA-ndhD*, *psbI-trnS-GCU*, *psbC-trnS-UGA*, *rpl32-trnS-AGA*, *rpl33-rps18*, *trnW-CCA-trnP-UGG*, and *psbZ-trnG-*

10.3389/fpls.2023.1070600

TABLE 2 Gene types and functional classification of the *Cerasus* chloroplast genome.

Category	Gene group	Gene symbol				
Self-replication	Ribosomal RNA genes	*rrn4.5*[a]	*rrn5*[a]	*rrn16*[a]	*rrn23*[a]	
	Transfer RNA genes	*trnA-UGC*[a;b]	*trnC-GCA*	*trnD-GUC*	*trnE-UUC*	*trnF-GAA*
		trnfM-CAU	*trnG-GCC*	*trnG-UCC*[b]	*trnH-GUG*	*trnI-CAU*[a]
		trnI-GAU[a;b]	*trnK-UUU*[b]	*trnL-CAA*[a;b]	*trnL-UAA*[b]	*trnL-UAG*
		trnM-CAU	*trnN-GUU*[a]	*trnP-UGG*	*trnQ-UUG*	*trnR-ACG*[a]
		trnR-UCU	*trnS-AGA*	*trnS-GCU*	*trnS-GGA*	*trnS-UGA*
		trnT-GGU	*trnT-UGU*	*trnV-GAC*[a]	*trnV-UAC*[b]	*trnW-CCA*
		trnY-GUA				
	Small subunit of ribosome	*rps2*	*rps3*	*rps4*	*rps7*[a]	*rps8*
		rps11	*rps12*[a;b]	*rps14*	*rps15*	*rps16*[b]
		rps18	*rps19*[d]	*rps36*		
	Large subunit of ribosome	*rpl2*[a;b]	*rpl14*	*rpl16*[b]	*rpl20*	*rpl22*
		rpl23[a]	*rpl32*	*rpl33*	*rpl36*	
	DNA-dependent RNA polymerase	*rpoA*	*rpoB*	*rpoC1*[b]	*rpoC2*	
Photosynthesis	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*
	Subunits of cytochrome	*petA*	*petB*[b]	*petD*[b]	*petG*	*petL*
		petN				
	Subunits of ATP synthase	*atpA*	*atpB*	*atpE*	*atpF*[b]	*atpH*
		atpI				
	Large subunit of RuBisCO	*rbcL*				
	Subunits of NADH dehydrogenase	*ndhA*[b]	*ndhB*[a;b]	*ndhC*	*ndhD*	*ndhE*
		ndhF	*ndhG*	*ndhH*	*ndhI*	*ndhJ*
		ndhK				
Other genes	Maturase	*matK*				
	Translational initiation factor	*infA*				
	Envelope membrane protein	*cemA*				
	Subunit of acetyl-CoA	*accD*				
	C-type cytochrome synthesis gene	*ccsA*				
	Protease	*clpP*[c]				
	Proteins of unknown function	*ycf1*[a;d]	*ycf2*[a]	*ycf4*	*ycf3*[c]	*ycf15*[a]

[a]Two gene copies in inverted repeats; [b]gene containing a single intron; [c]gene containing two introns; [d]gene divided into two independent transcription units.

GCC (*Pi* = 0.01334). Moreover, there were 14 IGRs with *Pi* > 0.01 and *S* > 10 (Table S2B; Figure S3). We further analyzed the sequence divergence patterns among all cp genomes. Finally, 506 single nucleotide polymorphisms (SNPs), 59 nucleotide substitution (NS) loci, and 1898 indel loci were identified through the nucleotide alignment (Table S3).

Comparing the IR/LSC and IR/SSC boundaries of the 11 cp genomes of *Cerasus* species revealed the contraction and expansion of IRs with minimal variation at the boundaries. The boundary region of IRA/SSC appears to be relatively stable (Figures 3, S3). That is, the boundary gene *ycf1* showed high conservation among *Cerasus* species with a length of 5606 bp for the majority of the

10.3389/fpls.2023.1070600

FIGURE 2
Chloroplast genome sequence comparison of 11 *Cerasus* species, apple, pear, and grape based on mVISTA. The similarity graphical information portrays sequence identity with *Prunus pseudocerasus* (NC030599) as reference. A cut-off of 50% identity is used for the plots. In each plot, the Y-axis represents percent identity (50–100%).

Cerasus species (7/11) analyzed. In *P. serrulata*, *P. tomentosa*, and *P. japonica*, *ycf1* had greater extension than found in the other species to different degrees (Figure 3). Both the *ycf1* pseudogene and *ndhF* gene were at the IRB/SSC borders, which partially overlapped in the cp genomes of *Cerasus* species. The IRB/LSC junction was largely located in *rps19*, close to *rpl22* and *rpl2*, and except in *P. avium*, extension of the LSC resulted in larger contraction of IRB toward the *rpl22* direction. Concerning the IRA/LSC boundary, the junction site was the *rps19* gene in all *Cerasus* species, except for *P. avium* in which *rps19* showed a contraction in IRA regions, being 48 bp away from the boundary. In addition, *trnH* was consistently observed in all plastomes, which was located 5–86 bp away from the border.

3.3 Repeat sequence analysis

The types and distribution of SSRs were analyzed in the cp genomes of the 11 *Cerasus* species. A total of 634 SSRs were identified by MISA, 61.51% of which were distributed in the IGR and 133 of which were found in CDSs (Figure 4A; Table S4A). Except for *P. tianshanica*, the SSRs were mainly enriched in the IGRs and CDSs, both accounting for 41.67% of all SSRs. For the other *Cerasus* species, SSRs were more abundant in IGRs than in other regions (Figure 4A; Table S4A). Moreover, four types of SSRs were identified: mono-, di-, and tri-nucleotide, and complex repeats. Mononucleotide repeats were the most common, accounting for 85.65% of the total (n =543, range 45–54; Table

10.3389/fpls.2023.1070600

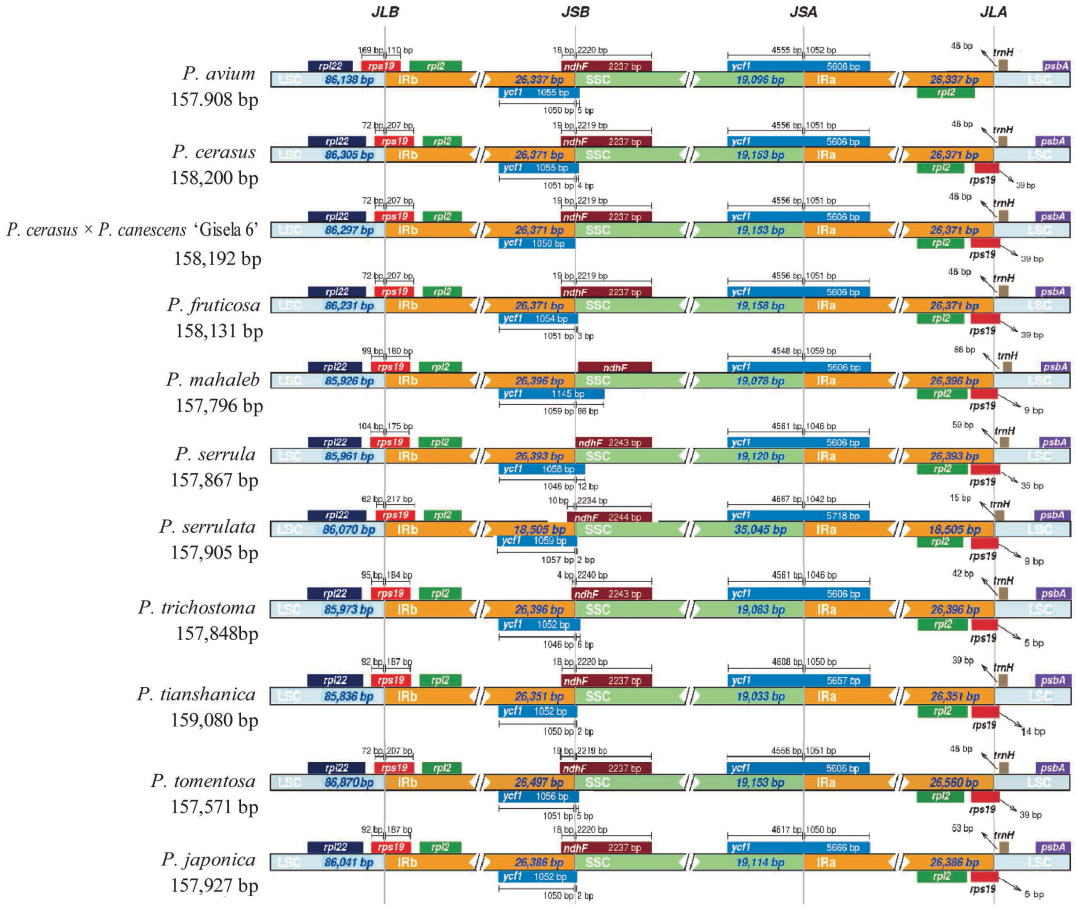


FIGURE 3
Comparison of the inverted repeat (IR)/large single copy (LSC) boundaries and IR/short single copy (SSC) boundaries among chloroplast genomes of 11 *Cerasus* species. JLA and JLB indicate the junction sites between the SSC and the two IRs (IRA and IRB); JSA and JSB denote the junction sites between the SSC and the two IRs.

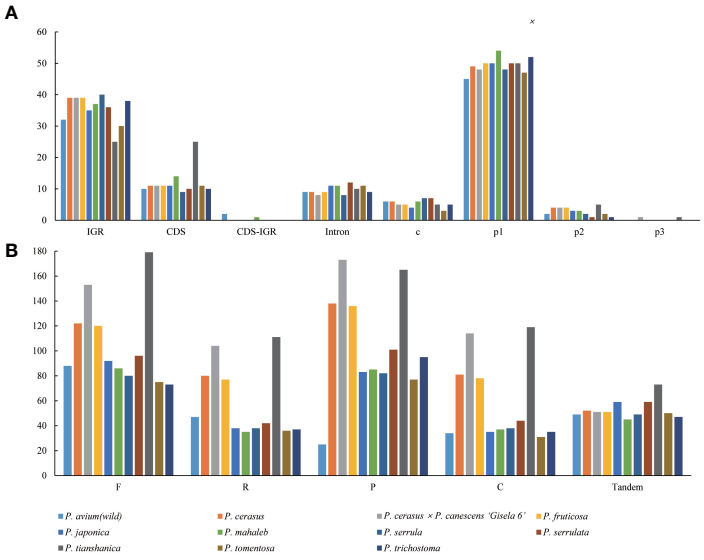


FIGURE 4
Repetitive motif abundance in 11 *Cerasus* species computed by REPuter and Tandem Repeats Finder. **(A)** Distribution and types of simple sequence repeats in the eleven chloroplast genomes. **(B)** Number of repeat types in the eleven chloroplast genomes.

S4A). By contrast, only 31 dinucleotide repeats were identified (ranging from 1 to 5), accounting for 4.89% of the total, and only two trinucleotide repeats were found, in *P. cerasus* × *P. canescens* 'Gisela 6' and *P. tianshanica*, respectively. No other polynucleotides were detected. In addition, the composition of the mononucleotide repeats was mostly A/T, with C and G repeats accounting for less than 6% of these repeats (Table S4A), and all dinucleotide repeats were composed of AT/TA. For the non-overlapping repeats, there were 1164 forward repeats, 645 reverse repeats, 1160 palindromic repeats, 646 complement repeats, and 585 tandem repeats identified using REPuter (Kurtz et al., 2001) and Tandem Repeats Finder for the *Cerasus* plastomes (Table S4B). Forward repeats were the most abundant (n = 73–179), followed by palindromic repeats (n = 25–173). Tandem repeats were the least abundant repeat type, ranging from 45 in *P. mahaleb* to 73 in *P. tianshanica* within 5–249 bp (Table S4). Dispersed repeats were more common in *P. cerasus* × *P. canescens* 'Gisela 6' and *P. tianshanica* than in the other species (Figure 4B and Table S4B).

3.4 Codon usage bias and gene selective pressure analysis

A total of 49 *Cerasus* genes were selected based on the 350-bp length threshold for codon bias identification. Codon usage measures such as Nc; silent T, C, A, and G at the 3rd codon position (T3s, C3s, A3s and G3s, respectively); total number of amino acids (L_aa); aromaticity; and grand average of hydropathicity (GRAVY) were estimated (Tables S4A, S5). The number of synonymous codons (L_sym) was approximately 49.7 and the range of L_aa, except for the termination codon (TER), was from 24,034 in *P. cerasus* × *P. canescens* 'Gisela 6' to 24,421 in *P. serrulata*, with a relative synonymous codon usage (RSCU) value ranging from 0.38 (CUC) to 1.96 (UUA) (Tables S5B, C; Figures S5, 5). The average Nc was 49.74 for CDSs of the 11 *Cerasus* species, and the T3s, C3s, A3s, and G3s was 0.4674, 0.1718, 0.4356, and 0.1817, respectively. The mean of GC3s was 0.269 and the GC content was 0.376 (Table S5A). In addition, leucine and isoleucine were the most common codons (Figure 5). Methionine (AUG) and tryptophan (UGG) were each encoded by only one codon, and showed no codon bias (Figure 5). Codon usage was biased toward A and T at the third codon position. Almost all A/U-ending codons had RSCU values larger than 1.0, except for Ile-AUA, Leu-CUA, and TER-UGA, whereas all C/G-ending codons had RSCU values ≤ 1, except for Leu-UUG (Figure S5; Table S5C).

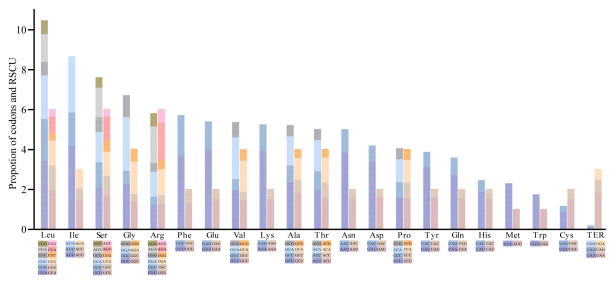


FIGURE 5
Codon content of 20 amino acids and termination codons in protein-coding genes of *Cerasus* plastomes. The left columns of the histogram present codon usage and the right columns denote the relative synonymous codon usage (RSCU) values. Colors correspond to codons listed underneath the columns.

The synonymous (Ks) and non-synonymous (Ka) nucleotide substitution patterns are very important markers in gene evolution studies (Kimura, 1989). In all protein-coding genes of *Cerasus*, *ndhB*, *petL*, *psbH*, *rpl33*, *rps18*, and *ycf15* had no nonsynonymous rate change, and 20 genes (*psaJ*, *psbA*, *atpE ndhE*, *ndhJ*, *psaC*, *petD*, *ycf3*, *ycf4*, three *rpl* genes, three *rps* genes, and five *psb* genes) had no synonymous rate change. Fourteen genes showed neither substitution, including *rpl23*, *rps7*, *rps8*, *rps12*, *petG*, *petN*, *psaI*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbN*, and *psbT*. Table S6 lists the genes with both Ka and Ka substitutions. A Ka/Ks ratio < 1, especially less than 0.5, indicates purifying selection; Ka/Ks > 1 indicates likely positive selection; and Ka/Ks values close to 1 show neutral evolution, or relaxed selection (Kimura, 1989; Ivanova et al., 2017). The average Ka/Ks ratio analyzed in the 11 genomes was 0.3158 for 34 protein-coding genes, which were not region-specific. Most protein-coding genes have undergone purifying selection: 33 coding genes showed Ka/Ks < 1, ranging from 0.0301 (*atpA*) to 0.8861 (*ccsA*). *ycf2*_IRA showed Ka/Ks > 1, whereas the ratio for the other *ycf2* gene on IRB was 0.2908. The *ccsA* and *ndh* genes (except for *ndhK* in the LSC region) were located in the SSC region, *ycf1* was located at the boundary of the LSC and IR regions, and other genes existed in the LSC. Moreover, Ka/Ks ratios in the range of 0.5–1 were observed for the genes *petA*, *rps15*, and *ccsA*, with the value of *ccsA* being close to 1, indicating neutral evolution (Table S6A; Figure S6). The Ka/Ks values of the remaining genes were between 0.04 and 0.50, with *rpoA*, *atpA*, *pet*B, *cemA*, and *rpoC1* showing patterns of strong purifying selection pressure (Ka/Ks < 0.1). For *ycf2*_IRA, *matK*, *rpoC2*, *ccsA*, and *ndhI*, the Ka/Ks ratio was greater than 1 in a few cases between the species (Table S6B), especially for *ycf2*_IRA in the comparison of *P. tianshanica* with other species. In addition, we analyzed the Ka/Ks ratios of the 10 cp genomes for comparison with *P. avium*, which reflected the selection pressure for the *P. avium* cp genome (Figure S7). The average Ka/Ks values of most genes were less than 0.5, except for *matK*, *rpoC2*, and *ycf2*_IRA. The Ka/Ks value of *matK* was greater than 1 in the comparisons of *P. avium* with *P. japonica*, *P. tomentosa*, and *P. tianshanica* (Table S6C).

We found that photosynthesis genes had varying Ka/Ks ratios, which were all less than 1, including one large subunit of the RuBisCO gene *rbcL* (0.1799–0.4581); two subunits of the photosystem II genes *psbB* and *psbC* (0.0643–0.2921); two subunits of the cytochrome genes *petA* and *petB* (0.0561–0.6536); and three subunits of the ATP synthase genes *atpA*, *atpB*, and *atpF* (0.0301–0.3146). The ratios of five subunits of NADH dehydrogenase genes (*ndhA*, *ndhD*, *ndhH*, *ndhG*, *ndhG* and *ndhI*) ranged from 0.04701 to 1.7677, and the Ka/Ks ratio of *ndhI* was greater than 1. Ka/Ks ratios of self-replicating genes were as follows: 0.0566–0.8818 for ribosomal protein small subunit genes (*rps3*, *rps4*, *rps15*, *rps16*), 0.2447–0.496 for ribosomal protein large subunit genes (*rpl16* and *rpl20*), and 0.0327–1.1722 for DNA-dependent RNA polymerase genes (*rpoA*, *rpoB*, *rpoC1*, and *rpoC2*). Among these, only *rpoC2* had Ka/Ks > 1. Among the other genes, *matK*, *ccsA*, *cemA*, *clpP*, *accD*, *ycf1*, and *ycf2*, the Ka/Ks ratios of *matK*, *ycf2*, and *ccsA* were more than 1 (Table S6B).

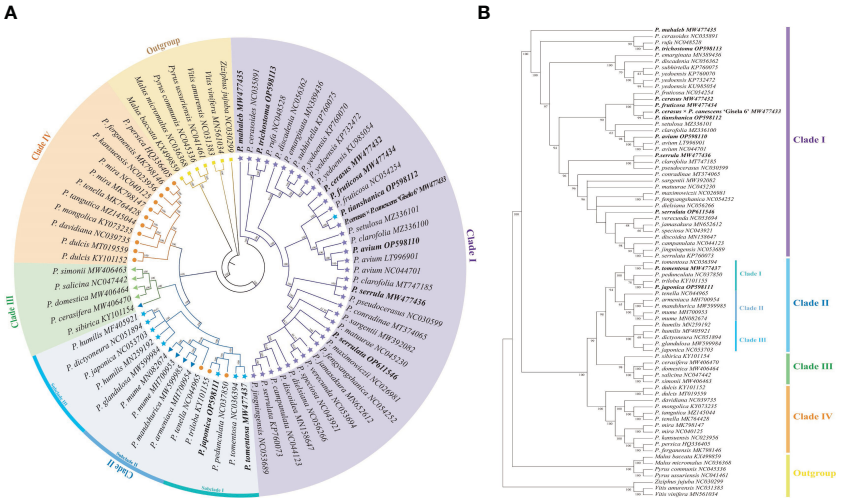


FIGURE 6
Phylogenetic tree reconstruction of 45 *Cerasus* species and 23 other Prunoideae species based on whole chloroplast genome sequences. **(A)** Phylogenetic tree reconstruction using the program Mrbayes of Geneious Prime v2022.0.2. Numbers above the lines represent the Bayesian inference posterior probability (percent). **(B)** Phylogenetic tree reconstruction using Maximum likelihood method of MEGA v11. Numbers below the lines represent the bootstrap support values.

3.5 Phylogenetic analysis

Organelle genome sequencing plays a key role in deciphering the evolutionary phylogenomics and cladistics of plant species. Phylogenetic relationships of *Cerasus* species were estimated with 6 datasets respectively, using Bayesian inference (BI) and ML methods. The results showed the tree topologies based on LSC, SSC, CDS and IGS datasets is basically consistent with complete plastome, especially using BI method (Figures S8–S12). Phylogenetic relationships were almost as consistent with BI and ML with the complete cp genome sequences (Figure 6). In the ingroup, the subgenera *Cerasus*, *Prunus*, *Armeniaca*, and *Amygdalus* were divided into four major clades (clade I, clade II, clade III, and clade IV) with BI posterior probabilities (BIPP) of 100% and ML bootstrap support (MLBS) of 99–100%. All of the clade I species belonged to subgenus *Cerasus* and all clade IV species belonged to subgenus *Amygdalus*. In clade III, only *P. sibirica* was an *Armeniaca* species, whereas the others belonged to *Prunus*. Among the four clades, clade II showed a relatively more complex composition, including dwarf cherry species (*P. tomentosa*, *P. japonica*, *P. glandulosa*, *P. humilis*, and *P. dictyoneura*) of *Cerasus*, *Amygdalus* (*P. tenella*, *P. pedunculata*, *P. triloba*), and *Armeniaca* (*P. mume*, *P. mandshurica*, and *P. armeniaca*). Through rebuilding the phylogenetic tree, clade II could be further divided into three small clades (BIPP = 99.87–100%; MLBS = 94–100%), named subclade I, subclade II, and subclade III, respectively. All species in subclade III belonged to *Cerasus*, and subclade I was also dominated by *Cerasus* except for *P. triloba*. Conversely, most of the subclade II species belonged to the subgenus *Armeniaca*, except for *P. tenella*. In clade I, *P. mahaleb* formed a separate branch (Figure 6). *P. trichostoma* was the closest relative to *P. rufa*. *P. cerasodies*, *P. fruticosa*, *P. cerasus*, *P. tianshanica*, and *P. cerasus* × *P. canescens* 'Gisela 6' were divided into the same small clade, which was close to the clade formed by *P. avium*, *P. clarofolia*, and *P. setulasa*. *P. serrula* was divided into a clade with *P. pseudocerasus* and *P. clarofolia*. *P. serrula* was the closest relative to *P. fengyangshanica*, *P. jamasakura*, *P. dielsiana*, and seven other species. Surprisingly, *P. tomentosa* and *P. japonica* were separated from the majority of *Cerasus* species and were most closely related to *P. triloba* and *P. pedunculata* of the subgenus *Amygdalus* (Figure 6).

4 Discussion

4.1 *Cerasus* cp genome features and genomic variation

The cp genome exhibits maternal inheritance in contrast to the nuclear genome (Palmer, 1985). Because of its evolutionarily conserved structure, sequence length, and constituent genes, the cp genome has been widely used in analyses of genetic variation and phylogeny with moderate base replacement (Palmer, 1985; Ravi et al., 2007; Wicke et al., 2011). In this study, the 11 *Cerasus* cp genomes presented a typical quadripartite structure (LSC, SSC, IRA, and IRB), as reported for other *Prunus* species (Xue et al., 2019; Zhang et al., 2021; Li et al., 2022) and land plants (Wicke et al., 2011; Zhang et al., 2017; Liu et al., 2020). The *Cerasus* cp genome size (157,571–158,830 bp) was similar to that of previously reported *Prunus* species such as *P. pseudocerasus* (157,834 bp), *P. dielsiana* (158,005 bp), *P. clarofolia* (157,899 bp), *P. mira* (158,153 bp), and *P. salicina* (157,916 bp) (Feng et al., 2018; Bao et al., 2019; Xue et al., 2019; Zhao et al., 2019; Li et al., 2022). This indicated that the length of the cp sequence is relatively conserved with only moderate variation among species of the *Prunus* genus.

Comparative analysis of *Cerasus* species showed that the LSC and SSC regions were more divergent than the IR regions (Figure 2), whereas the CNSs showed significant divergence (Figure S3), consistent with findings for other species (Xue et al., 2019; Liu et al., 2020; Wang et al., 2021). One of the most important factors contributing to the

variation in plastome size between species may be the expansion and contraction of IR boundaries (Xue et al., 2019; Zhang et al., 2021). The contraction and expansion of *rps19* and *ndhF* were predicted as the main contributors to the overall variation observed among the *Cerasus* cp genomes, followed by expansion of *ycf1* toward the SSC region (Figures 3, S4). The three genes at IR boundaries were consistent in *P. mume*, *P. salicina*, and *P. armeniaca* (Xue et al., 2019). Significant IR contraction of *rps19* and *ndhF* was observed in the plastomes of some *Cerasus* and other *Prunus* species (Zhang et al., 2021; Wang et al., 2022). In addition, displacement of the *trnH* gene at the IR/LSC locus was detected in the aforementioned studies and the current study. This pattern of these four genes located at IR boundaries was also found in species of other genera, including *Malus* (Zhang et al., 2021; Wang et al., 2022), *Diospyros* (Heinze et al., 2016), and *Morella* (Liu et al., 2017). Therefore, although the IR regions are highly conserved for stabilizing the cp genome structure (Figure 2) (Marechal and Brisson, 2010), some changes (especially in *rps19, ndhF*, and *ycf1*) are evident at the IR border areas among *Cerasus* species, in line with reports for other species and genera.

We annotated 115 unique genes in this study (Table 2), which is similar to the findings reported for other *Prunus* plastomes with 110–115 unique genes (Xue et al., 2019; Wang et al., 2021; Zhang et al., 2021; Li et al., 2022). There were four rRNA genes detected (4.5S rRNA, 5S rRNA, 16S rRNA, and 23S rRNA), which coincides with reports for land plants (Sugiura, 1992). The differences were mainly reflected in tRNA and protein-coding genes, such as *ycf* genes (Xue et al., 2019; Zhang et al., 2021). The *ycf15* gene was detected in all 11 *Cerasus* cp genomes, but is lacking in some *Prunus* (*sensu lato*) plastomes (Zhang et al., 2021). According to previous reports, angiosperm plastomes harbor approximately 70–88 protein-coding genes (Liu et al., 2020) and 80 unique protein-coding genes were annotated in this study. We found some CDSs and CNSs with relatively high nucleotide diversity (Table S2; Figure S3), which was in line with previous research (Xue et al., 2019; Liu et al., 2020; Wang et al., 2021). Analyses of cp genes and genomes have largely contributed to resolving portions of the plant tree of life (Ravi et al., 2007). Various genes and IGRs have been identified as evolutionarily significant markers, which have been widely used for phylogenetic analyses (Ravi et al., 2007). Intergenic spacer regions were proposed to be the best barcoding candidates (Ravi et al., 2007), which was also confirmed in this study, and we further found that the *Pi* of the IGRs was higher than that of the CDSs (Table S2; Figure S3). Nevertheless, both the IGRs and CDSs can serve as useful molecular markers. Some genes (*matK*, *rps16*, *rbcL*, *rpl16*, *ndhA*, *ndhF*, and *ycf1*) and IGRs (*rps16-trnQ*, *petN-psbM*, *rps15-ycf1*, *trnL-trnF*) exhibited high nucleotide diversity (Table S2), which have also been used for phylogenetic and phylogeography analyses (Ravi et al., 2007; Chavez et al., 2016; Uchoi et al., 2016; Yang et al., 2016; Khan et al., 2018).

4.2 Evolutionary and phylogenetic analysis

Repeats play an important role in genome rearrangement, which can increase the probability of replication fork stagnation, causing an error to recruit persistently specific sequence regions over evolutionary time scales (Mcdonald et al., 2011). Repetitive sequences may facilitate intermolecular recombination and enhance plastome diversity, as an abundance of sequence repeats results in genome regions with increased sequence diversity in prokaryotes and eukaryotes (Mcdonald et al., 2011; Liu et al., 2020). We also found abundant repeats in the *Cerasus* genome, including dispersed, palindromic, and tandem repeats, along with SSRs (Figure 4A; Table S4A). More abundant SSRs were detected in IGRs for most species, except for *P. tianshanica* (Figure 4A; Table S4A) and other *Prunus* species (Wang et al., 2021; Zhang et al., 2021). Combined with the visualization of aligned genome sequences, CNSs showed more significant divergence (Figure 2). This suggested that repetitive CNSs might be the main force promoting cp genome rearrangement in *Prunus* species (Wang et al., 2021). Furthermore, given the characteristics of maternal inheritance, conservation, and simple structure of the cp genome, cp microsatellites with a high degree of polymorphism can serve as useful molecular markers to identify genetic relationships, population genetic structure, and phylogeography patterns at the inter- and intrapopulation levels (Decroocq et al., 2004; Wang et al., 2021). We only detected mono-, di-, and tri-nucleotide repeats in the *Cerasus* cp genomes (Figure 4; Table S4) with a greater content of A/T repeats than of G/C repeats, similar to the results of other studies in the genus *Prunus* (Xue et al., 2019; Zhang et al., 2021). The SSRs identified in this study provide useful information for developing genetic markers to further study the population genetics, evolution, and breeding of the subgenus *Cerasus*, as well as for the identification and conservation of *Cerasus* species. Repetitive sequences are also essential for research on indels and substitutions (Wang et al., 2021), which are highly abundant in the plastome of *Cerasus* (Table S3) and other members of the family Rosaceae (Zhang et al., 2021).

Codon usage and synonymous/nonsynonymous substitutions play an important role in cp plastome evolution (Ivanova et al., 2017; Huang et al., 2021). Mutation is one of the most essential factors affecting codon usage, thus influencing the evolutionary course (Ivanova et al., 2017; Wang et al., 2021). Moreover, codon-choice patterns are considered to be highly conserved during the evolution process (Ikemura, 1985). An RSCU value > 1, < 1, or = 1 indicates preference, low usage, and no preference for a codon, respectively (Sharp and Li, 1987). We found biased codon usage in the *Cerasus* cp genome, with 19 amino acids having an RSCU value >1 (Figure S5). The codon profile showed strong bias toward the use of A/T in the third-base position, which appears to be a general phenomenon (Murray et al., 1989). Leucine and isoleucine appeared the most frequently and were biased toward UUA and AUU, respectively (Figures S5, 5), whereas cysteine was the least frequently detected, with the third base also biased toward T (UGU) in codon usage (Figures S5, 5). Among the three stop codons, there was clear usage bias toward UAA (RSCU > 1.00) (Figure S5). These results are largely consistent with reports of the cp genomes in *P. zhengheensis* (Huang et al., 2021; Wang et al., 2021) and other species (Alzahrani et al., 2020; Liu et al., 2020). Two amino acids, methionine (AUG) and tryptophan (UGG), showed no codon usage

bias (RSCU = 1.00) (Figure S5). In other words, all amino acids are encoded by 2–6 synonymous codons with the exception of methionine and tryptophan (Murray et al., 1989).

The Ka and Ks nucleotide substitution rates as well as the Ka/Ks ratio are widely used to evaluate the sequence divergence and purifying selection in protein-coding genes. In most genes, with the exception of very rapidly evolving genes, Ka nucleotide substitutions occur less frequently than Ks owing to the action of purifying selection (Ivanova et al., 2017). Ks nucleotide substitutions generally occur more frequently than Ka substitutions (Liu et al., 2017), which was also detected for *Cerasus* in this study. Among the changed genes, almost all Ka/Ks ratios were less than 1.0 (Tables S5, S6), providing evidence for purifying selection on the protein-coding genes of the cp genome in the genus *Prunus* and family Rosaceae. According to the Ka/Ks values, we found that *rpoA*, *atpA, petB*, *cemA*, and *rpoC1* exhibited strong purifying selection. The *ccsA* was under neutral selection, whereas *ycf2* genes showed a signature of possible positive selection during the course of *Cerasus* evolution (Table S6), owing to an inserted fragment in the middle of the gene (Figure S2A), in line with the Ka/Ks analyses. However, these patterns are in contrast to previous research on *Cerasus* (Zhang et al., 2021), in which the *matK* and *rpoC2* genes both showed signatures of positive selection, along with other genes such as *ndhF*, *atpA*, and *psaA*. Under conditions of extreme temperature and changing light intensity, the *ndhF* gene can balance the redox levels to maintain or enhance photosynthetic performance (Martin et al., 2009; Zhang et al., 2021). In addition, Zhang et al. (2021) detected strong signatures of positive selection in several genes of Rosaceae, including *rpoA*, *rps16, rps18, psaA, psbL, rbcL, ndhD, ndhF, accD, ycf1*, and *ycf2*. In particular, the *rbcL* gene encodes the large subunit of RuBisCO, which is one of the most useful enzymes for studying plant evolution, serving as a model protein in several studies owing to its response to environmental pressure and climate shifts (Hermida-Carrera et al., 2017; Zhang et al., 2021). In addition to *rbcL*, *ndhF*, *ycf1*, and *rps18* had high *Pi* values in this study, which can help Rosaceae woody fruit trees efficiently capture light energy to obtain sufficient nutrition for growth and development as an adaptation under extreme and variable environmental conditions (Table S2). Hence, it is necessary to further study the patterns of synonymous and non-synonymous substitutions among *Cerasus* species, which can provide new insight into the evolution of Rosaceae.

The subgenus *Cerasus* can be classified in two sections of true cherry (*Cerasus sensu stricto*) and dwarf cherry (*Microcerasus*), according to Yü et al. (1986). The genetic evolution of *Cerasus* species has been a long-standing open research question (Liang et al., 2018; Zhang et al., 2021). Given its maternal inheritance, the cp genome has been widely used for species classification and evolutionary analyses (Wang et al., 2022). Therefore, we reconstructed the *Cerasus* phylogenetic tree based on the complete cp genome (Figure 6). In line with the classification proposed by Yü et al. (1986) and others, true cherry species were classified in a single group as clade I (Figure 6). Surprisingly, *P. tianshanica*, as a *Microcerasus* species, also clustered in the true cherry group and formed a clade with *P. cerasus*, *P. fruticosa*, *P. cerasus* × *P. canescens* 'Gisela 6', and *P. avium*. Moreover, *P. mahaleb* in clade I was separated from other true cherry species, which is in line with previous reports (Chin et al., 2014; Zhang et al., 2021). Further subdivisions of true cherry match the existing classification (Yü et al., 1986; Webster and Looney, 1996) proposing *P. mahaleb* or *P. emarginata*, *P. pennsylvanica*, and *P. prunifolia* as a separate group, named section *Mahaleb* Focke (Table S7). However, in the present study, *P. emarginata* was not grouped with *P. mahaleb* (Figure 6). In addition, we found that some true cherry species did not follow the grouping of existing classifications (Yü et al., 1986; Webster and Looney, 1996) or varied to a certain extent at the section level, such as *P. tianshanica*, *P. serrula*, and *P. serrulata* (Figure 6; Table S7). *Microcerasus* also showed a significant difference from other taxa (Table S7) (Yü et al., 1986; Webster and Looney, 1996), which were grouped with *Amygdalus, Armeniaca*, and *Prunus* as a subclade in this study (Figure 6). These relationships were also reported in previous studies (Chen et al., 2018; Zhang et al., 2021; Li et al., 2022). *Microcerasus* showed close evolutionary relationships to *Amygdalus* (*P. tenella*, *P. pedunculata*, and *P. triloba*) and *Armeniaca* (*P. mume*, *P. mandshurica*, and *P. armeniaca*) species. Specifically, *P. japonica* and *P. tomentosa* were the closest relatives to *Amygdalus* (*P. triloba* and *P. pedunculata*). Previous studies also revealed that *P. pesica* in *Amygdalus* was closely related to *Microcerasus* (Chen et al., 2018; Wang et al., 2020; Zhang et al., 2021). Hence, during the evolution process, true cherry formed a distinct group, while *Microcerasus* remained genetically closer to *Amygdalus*, *Armeniaca*, and *Prunus* (*sensu stricto*) than to true cherry. These results can be supported by multi-cp genome comparative (Zhang et al., 2021) and whole-genome analyses (Wang et al., 2020). Nevertheless, further breakdown of *Cerasus* (*sensu stricto*) should not be ignored based on these results, which contrast with the existing classification criteria (Yü et al., 1986; Webster and Looney, 1996). Accordingly, further study of the taxa of the subgenus *Cerasus*, especially *Microcerasus*, is necessary to enable breeding novel cherry cultivars and to gain a better understanding of the evolution of *Prunus sensu lato* specifically and Rosaceae plants more broadly.

5 Conclusion

Comparative analysis of the cp genome is a key approach to study the molecular evolution and reconstruct the phylogenetic tree of *Cerasus* species. The present analysis of the cp genomes of 11 *Cerasus* species showed that IR regions are more strongly conserved than the LSC and SSC regions, whereas the non-coding sequences showed more significant divergence than the coding regions. The contraction/expansion of *rps19* and *ndhF* at the IR boundaries were the main contributors to the observed variation among *Cerasus* cp genomes, as well as variation in *ycf1* and *trnH*. We identified 26 genes and IGRs with variations that can be used as potential molecular markers and candidate DNA barcodes for studying the phylogeny and phylogeography of cherry species. We further provided important evidence that *P. mahaleb* forms a unique clade among true cherry (*Cerasus sensu stricto*) due to plastid genome rearrangement. *Microcerasus* was found to be genetically closer to *Amygdalus*, *Armeniaca*, and *Prunus* (*sensu stricto*) than to true cherry species.

Moreover, *P. tianshanica* emerged as a noteworthy species given its close relationship to *P. avium*. Overall, these findings provide new insight and resources to breed novel cultivated sweet cherry and cherry rootstock in the future.

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

TW and Y-LC designed the research. TW performed the experiments and analyzed the data. Y-LC identified the plant materials and revised the manuscript. B-XQ, JZ, and K-SS revised the manuscript. FA assembled the sequences. TW, B-XQ, JZ, L-YP, and FA annotated the plastomes. TW, K-SS, L-YP, X-SH, TL, and P-KL collected the plant materials. FA and L-YP provided analysis support. All authors contributed to the article and approved the submitted version.

Funding

This work was financially supported by the Bingtuan Science and Technology Program (Grant number: 2021AB017) and the Agricultural Science and Technology Innovation and Transformation Project of Shaanxi province (Grant number: NYKJ-2022-YL(XN)47). The funders had no role in study design, data collection, data analysis, data interpretation, the writing of the manuscript, or decision to publish.

Acknowledgments

We are grateful to Associate Professor Zengqiang Qian from College of Life Sciences, Shaanxi Normal University, and Guoqing Bai from Xi'an Botanical Garden of Shaanxi Province (Institute of Botany of Shaanxi Province Shaanxi) and Engineering Research Centre for Conservation and Utilization of Botanical Resources for their useful advice and help in data analyses.

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.2023.1070600/full#supplementary-material

References

Alzahrani, D. A., Yaradua, S. S., Albokhari, E. J., and Abba, A. (2020). Complete chloroplast genome sequence of *Barleria prionitis*, comparative chloroplast genomics and phylogenetic relationships among acanthoideae. *BMC Genomics* 21, 393. doi: 10.1186/s12864-020-06798-2

Amiryousefi, A., Hyvonen, 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

Bao, W., Ao, D., Wuyun, T., Li, T., Wang, L., and Liu, H. (2019). The complete chloroplast genome of *Prunus mira* koehne (Prunoideae, rosaceae), a wild and indigenous peach on Tibet, China. *Mitochond. DNA B* 4, 3731–3733. doi: 10.1080/23802359.2019.1679048

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

Bortiri, E., Oh, S. H., Jiang, J., Baggett, S., Granger, A., Weeks, C., et al. (2001). Phylogeny and systematics of *Prunus* (Rosaceae) as determined by sequence analysis of ITS and the chloroplast trnL-trnF spacer DNA. *Syst. Bot.* 26, 797–807. doi: 10.1043/0363-6445-26.4.797

Brudno, M., Malde, S., Poliakov, A., Do, C. B., Couronne, O., Dubchak, I., et al. (2003). Glocal alignment: finding rearrangements during alignment. *Bioinformatics* 19 Suppl 1, i54–i62. doi: 10.1093/bioinformatics/btg1005

Chan, P. P., Lin, B. Y., Mak, A. J., and Lowe, T. M. (2021). tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Resour.* 49, 9077–9096. doi: 10.1093/nar/gkab688

Chavez, D. J., Beckman, T. G., and Chaparro, J. X. (2016). Identifying the north American plum species phylogenetic signal using nuclear, mitochondrial, and chloroplast DNA markers. *J. Am. Soc Hortic. Sci.* 141, 623–644. doi: 10.21273/jashs03875-16

Chen, T., Wang, Y., Wang, L., Chen, Q., Zhang, J., Tang, H. R., et al. (2018). The complete chloroplast genome of tomentosa cherry *Prunus tomentosa* (Prunoideae, rosaceae). *Mitochond. DNA B* 3, 672–673. doi: 10.1080/23802359.2018.1476068

Chin, S. W., Shaw, J., Haberle, R., Wen, J., and Potter, D. (2014). Diversification of almonds, peaches, plums and cherries - molecular systematics and biogeographic history of *Prunus* (Rosaceae). *Mol. Phylogen. Evol.* 76, 34–48. doi: 10.1016/j.ympev.2014.02.024

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

Decroocq, V., Hagen, L. S., Favé, M.-G., Eyquard, J.-P., and Pierronnet, A. (2004). Microsatellite markers in the hexaploid *Prunus domestica* species and parentage lineage of three European plum cultivars using nuclear and chloroplast simple-sequence repeats. *Mol. Breed.* 13, 135–142. doi: 10.1023/B:MOLB.0000018761.04559.b3

Dunning, L. T., Olofsson, J. K., Papadopulos, A. S. T., Hibdige, S. G. S., Hidalgo, O., Leitch, I. J., et al. (2022). Hybridisation and chloroplast capture between distinct themeda triandra lineages in Australia. *Mol. Ecol* 31, 5846–5860. doi: 10.1111/mec.16691

Feng, Y., Liu, T., Wang, X. Y., Li, B. B., Liang, C. L., and Cai1, Y. L. (2018). Characterization of the complete chloroplast genome of the Chinese cherry *Prunus*

pseudocerasus (Rosaceae). *Conserv. Genet. Resour.* 10, 85–88. doi: 10.1007/s12686-017-0770-9

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, W59–W64. doi: 10.1093/nar/gkz238

Greiner, S., Rauwolf, U., Meurer, J., and Herrmann, R. G. (2011). The role of plastids in plant speciation. *Mol. Ecol.* 20, 671–691. doi: 10.1111/j.1365-294X.2010.04984.x

Hahn, C., Bachmann, L., and Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads-a baiting and iterative mapping approach. *Nucleic Acids Res.* 41, e129. doi: 10.1093/nar/gkt371

Heinze, B., Fu, J., Liu, H., Hu, J., Liang, Y., Liang, J., et al. (2016). Five complete chloroplast genome sequences from *Diospyros*: Genome organization and comparative analysis. *PloS One* 11, e0159566. doi: 10.1371/journal.pone.0159566

Hermida-Carrera, C., Fares, M. A., Fernandez, A., Gil-Pelegrin, E., Kapralov, M. V., Mir, A., et al. (2017). Positively selected amino acid replacements within the RuBisCO enzyme of oak trees are associated with ecological adaptations. *PloS One* 12, e0183970. doi: 10.1371/journal.pone.0183970

Huang, X., Tan, W., Li, F., Liao, R., Guo, Z., Shi, T., et al. (2021). The chloroplast genome of *Prunus zhengheensis*: Genome comparative and phylogenetic relationships analysis. *Gene* 793, 145751. doi: 10.1016/j.gene.2021.145751

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

Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* 2, 13–34. doi: 10.1093/oxfordjournals.molbev.a040335

Ivanova, Z., Sablok, G., Daskalova, E., Zahmanova, G., Apostolova, E., Yahubyan, G., et al. (2017). Chloroplast genome Analysis of resurrection tertiary relict *Haberlea rhodopensis* highlights genes important for desiccation stress response. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.00204

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, G., Zhang, F., Gao, Q., Fu, P., Zhang, Y., and Chen, S. (2018). Spiroides shrubs on qinghai-Tibetan plateau: Multilocus phylogeography and palaeodistributional reconstruction of *Spiraea alpina* and *S. mongolica* (Rosaceae). *Mol. Phylogen. Evol.* 123, 137–148. doi: 10.1016/j.ympev.2018.02.009

Kimura, M. (1989). The neutral theory of molecular evolution and the world view of the neutralists. *Genome* 31, 24–31. doi: 10.1139/g89-009

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

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

Letunic, I., and Bork, P. (2021). Interactive tree of life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* 49, W293–W296. doi: 10.1093/nar/gkab301

Li, J., Yan, J., Yu, L., Bai, W., Nie, D., Xiong, Y., et al. (2022). The complete chloroplast genome of *Prunus clarofolia* (Rosaceae), a wild cherry endemic to China. *MITOCHONDRIAL DNA Part B* 7, 164–166. doi: 10.1080/23802359.2021.2016080

Liang, C., Wan, T., Xu, S., Li, B., Li, X., Feng, Y., et al. (2018). Molecular identification and genetic analysis of cherry cultivars using capillary electrophoresis with fluorescence-labeled SSR markers. *3 Biotech.* 8, 16. doi: 10.1007/s13205-017-1036-7

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

Liu, L.-X., Li, R., Worth, J. R. P., Li, X., Li, P., Cameron, K. M., et al. (2017). The complete chloroplast genome of Chinese bayberry *(Morella rubra*, myricaceae): Implications for understanding the evolution of fagales. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.00968

Marechal, A., and Brisson, N. (2010). Recombination and the maintenance of plant organelle genome stability. *New Phytol.* 186, 299–317. doi: 10.1111/j.1469-8137.2010.03195.x

Martin, M., Funk, H. T., Serrot, P. H., Poltnigg, P., and Sabater, B. (2009). Functional characterization of the thylakoid ndh complex phosphorylation by site-directed mutations in the *ndhF* gene. *Biochim. Biophys. Acta* 1787, 920–928. doi: 10.1016/j.bbabio.2009.03.001

Mcdonald, M. J., Wang, W. C., Huang, H. D., and Leu, J. Y. (2011). Clusters of nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences. *PloS Biol.* 9, e1000622. doi: 10.1371/journal.pbio.1000622

Moore, M. J., Soltis, P. S., Bell, C. D., Burleigh, J. G., and Soltis, D. E. (2010). Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proc. Natl. Acad. Sci. U.S.A* 107, 4623–4628. doi: 10.1073/pnas.0907801107

Morton, B. R. (1998). Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages. *J. Mol. Evol.* 46, 449–459. doi: 10.1007/PL00006325

Murray, M. G., and Thompson, W. F (1980). Rapid isolation of higher weight DNA. *Nucleic Acids Res* 8, 4321–4325. doi: 10.1093/nar/8.19.4321

Murray, E. E., Lotzer, J., and Eberle, M. (1989). Codon usage in plant genes. *Nucleic Acids Res.* 17, 477–498. doi: 10.1093/nar/17.2.477

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

Potter, D., Eriksson, T., Evans, R. C., Oh, S., Smedmark, J. E. E., Morgan, D. R., et al. (2007). Phylogeny and classification of rosaceae. *Plant Syst. Evol.* 266, 5–43. doi: 10.1007/s00606-007-0539-9

Ravi, V., Khurana, J. P., Tyagi, A. K., and Khurana, P. (2007). An update on chloroplast genomes. *Plant Syst. Evol.* 271, 101–122. doi: 10.1007/s00606-007-0608-0

Rozas, J., Ferrer-Mata, A., Sanchez-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

Sevindik, E., Murathan, Z. T., and Sevindik, M. (2020). Molecular genetic diversity of *Prunus armeniaca* l. (Rosaceae) genotypes by RAPD, ISSR-PCR, and chloroplast DNA (cpDNA) trnL-f sequences. *Int. J. Fruit S.* 20, S1652–S1661. doi: 10.1080/15538362.2020.1828223

Sharp, P. M., and Li, W.-H. (1987). The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential. *Nucleic Acids Res.* 15, 1281–1295. doi: 10.1093/nar/15.3.1281

Shaw, J., and Small, R. L. (2004). Addressing the "hardest puzzle in American pomology:" phylogeny of *Prunus* sect. prunocerasus (Rosaceae) based on seven noncoding chloroplast DNA regions. *Am. J. Bot.* 91, 985–996. doi: 10.3732/ajb.91.6.985

Sugiura, M. (1992). The chloroplast genome. *Plant Mol. Biol.* 19, 149–169. doi: 10.1007/bf00015612

Uchoi, A., Malik, S. K., Choudhary, R., Kumar, S., Rohini, M. R., Pal, D., et al. (2016). Inferring phylogenetic relationships of Indian citron (*Citrus medica* l.) based on *rbc*L and *mat*K sequences of chloroplast DNA. *Biochem. Genet.* 54, 249–269. doi: 10.1007/s10528-016-9716-2

Wang, L., Guo, Z., Shang, Q., Sa, W., and Wang, L. (2021). The complete chloroplast genome of *Prunus triloba* var. *plena* and comparative analysis of *Prunus* species: genome structure, sequence divergence, and phylogenetic analysis. *Braz. J. Bot.* 44, 85–95. doi: 10.1007/s40415-020-00685-6

Wang, T., Kuang, R.-P., Wang, X.-H., Liang, X.-L., Wanga, V. O., Liu, K.-M., et al. (2021). Complete chloroplast genome sequence of *Fortunella venosa* (Champ. ex benth.) C.C.Huang (Rutaceae): Comparative analysis, phylogenetic relationships, and robust support for its status as an independent species. *Forests* 12, 996. doi: 10.3390/f12080996

Wang, X., Wang, D., Gao, N., Han, Y., Wang, X., Shen, X., et al. (2022). Identification of the complete chloroplast genome of *Malus zhaojiaoensis* jiang and its comparison and evolutionary analysis with other *Malus* species. *Genes (Basel)* 13, 560. doi: 10.3390/genes13040560

Wang, P., Yi, S., Mu, X., Zhang, J., and Du, J. (2020). Chromosome-level genome assembly of *Cerasus humilis* using PacBio and Hi-c technologies. *Front. Genet.* 11. doi: 10.3389/fgene.2020.00956

Wang, D., Zhang, Y., 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, 77–80. doi: 10.1016/s1672-0229(10)60008-3

Webster, A. D., and Looney, N. E. (1996). *CHERRIES: Crop physiology, production and uses* (Canbridge: University Press).

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

Xue, C., Geng, F. D., Li, J. J., Zhang, D. Q., Gao, F., Huang, L., et al. (2021). Divergence in the *Aquilegia ecalcarata* complex is correlated with geography and climate oscillations: Evidence from plastid genome data. *Mol. Ecol.* 30, 5796–5813. doi: 10.1111/mec.16151

Xue, S., Shi, T., Luo, W., Ni, X., Iqbal, S., Ni, Z., et al. (2019). Comparative analysis of the complete chloroplast genome among *Prunus mume*, *P. armeniaca*, and *P. salicina*. *Hortic. Res.* 6, 89. doi: 10.1038/s41438-019-0171-1

Yang, J., Di, X., Meng, X., Feng, L., Liu, Z., and Zhao, G. (2016). Phylogeography and evolution of two closely related oak species (*Quercus*) from north and northeast China. *Tree Genet. Genom.* 12, 89. doi: 10.1007/s11295-016-1044-5

Yü, D. J., Lu, L. T., Ku, T. C., Li, C. L., and Chen, S. X. (1986). *Flora of China* (Beijing: Science Press).

Zarei, A., Erfani-Moghadam, J., and Mozaffari, M. (2017). Phylogenetic analysis among some pome fruit trees of rosaceae family using RAPD markers. *Biotechnol. Biotec. Eq.* 31, 289–298. doi: 10.1080/13102818.2016.1276414

Zhang, J., Wang, Y., Chen, T., Chen, Q., Wang, L., Liu, Z. S., et al. (2021). Evolution of rosaceae plastomes highlights unique *Cerasus* diversification and independent origins of fruiting cherry. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.736053

Zhang, X., Zhou, T., Kanwal, N., Zhao, Y., Bai, G., and Zhao, G. (2017). Completion of eight *Gynostemma* BL. (Cucurbitaceae) chloroplast genomes: Characterization, comparative analysis, and phylogenetic relationships. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.01583

Zhao, K., Zhou, Y., Zheng, Y., Chen, B., and Ziling, W. (2019). The chloroplast genome of *Prunus dielsiana* (Rosaceae). *Mitochond. DNA B* 4, 4033–4034. doi: 10.1080/23802359.2019.1688723

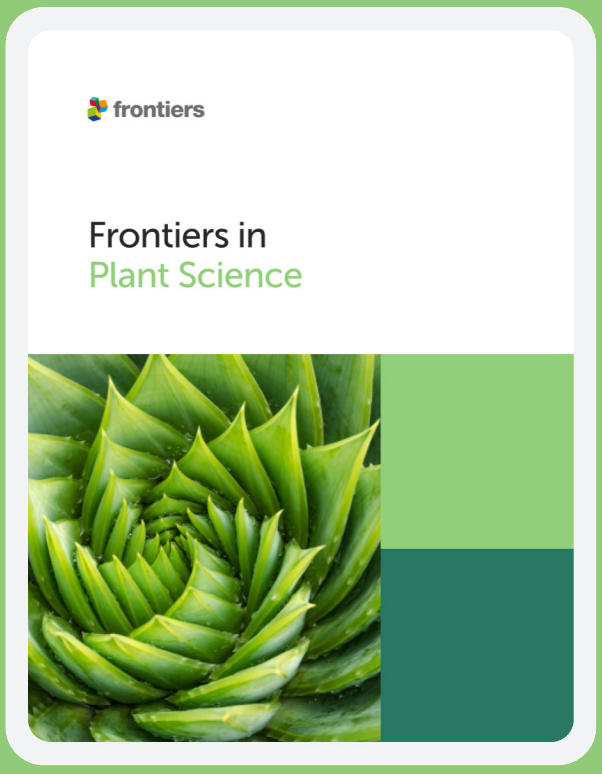
Frontiers in Plant Science

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