Brazilian Journal of Botany (2024) 47:93–103
https://doi.org/10.1007/s40415-023-00960-2

GENETICS & EVOLUTIONARY BIOLOGY - ORIGINAL ARTICLE

Comparative analysis of the complete chloroplast genome sequences of four *camellia* species

Bingqing Hao[1] · Yingying Xia[1] · Zhaoyuan Zhang[1] · Dongxue Wang[1] · Hang Ye[1] · Jinlin Ma[1]

Received: 19 April 2023 / Revised: 26 September 2023 / Accepted: 6 November 2023 / Published online: 26 December 2023

Abstract

Researching the photosynthetic characteristics based on the whole chloroplast genome sequence of *Camellia osmantha cv* '*yidan*' is important for improving production. We sequenced and analyzed the chloroplast (cp) genomes of *C. osmantha cv* '*yidan*'. The total cp genome length was 156,981 bp. The cp genomes included 134 genes encoding 81 proteins, 39 transfer RNAs, 8 ribosomal RNAs, and 6 genes with unknown functions. In total, 50 repeat sequences were identified in *C. osmantha cv* '*yidan*' cp genomes. Phylogenetic analysis showed that *C. osmantha cv* '*yidan*' is more closely related to *Camellia vietnamensis cv* '*hongguo*' and *Camellia oleifera cv* '*cenruan 3*' than to *Camellia semiserrata cv* '*hongyu 1*'. Our complete assembly of four Camellia cp genomes may contribute to breeding for high oil content plants and further biological discoveries. The results of this study provide a basis for the assembly of the entire chloroplast genome of *C. osmantha* cv 'yidan'.

Keywords *Camellia osmantha cv* '*yidan*' · Chloroplast genome · Genetic structure · Phylogenetic analysis

1 Introduction

The genus *Camellia*, which is used worldwide as an ornamental plant and for tea, belongs to the family Theaceae (Vijayan et al. 2012; Yang et al. 2013; Huang et al. 2014). *Camellia* oil is less known worldwide despite its use in China as an edible oil, as well as in Japan. *Camellia* is one of the four main oil-bearing trees in the world, in addition to palm, olive, and coconut (Robards et al. 2009).

Through years of research and experimentation, Guangxi Forestry Research Institute(GFRI) discovered the new species *C. osmantha* (Ma et al. 2012a, b). *C. osmantha* is easy to plant, grows rapidly, and has strong cold, heat, and drought tolerance (Ma et al. 2013; Liu et al. 2013) as well as high oil yield (Wang et al. 2014). *C. osmantha cv* '*yidan*' is recognized as a new variety of *C. osmantha* (Ma 2020). The plant height and crown width of 6-year-old *C. osmantha cv* '*yidan*' was 5.39 m and 7.17 m, respectively, and the oil production of a 5-year-old plant was 0.0590 kg·m^{-2} (Liang et al. 2017), almost double the standard oil yield for *C. oleifera* cultivars (0.0325 kg·m^{-2}). *Camellia* oil is also known as ''eastern olive oil'' because of the similarities in the chemical composition of *Camellia* and olive oils, with high amounts of oleic acid and linoleic acid, as well as low levels of saturated fats. At present, the total area of *C. osmantha cv* '*yidan*' production is over 1500 ha, mainly in Qinzhou, Laibin, Yulin, Yunnan, and Hainan, China.

In China, the planting area of *C. oleifera* reaches 4,466,700 ha, and the oil production is 600,000 tons. *Camellia* oil production needs to be further developed. *C. osmantha cv* '*yidan*' is a promising new species that produces 1590 kg of oil per hectare, doubling the standard oil productivity rate for *C. oleifera cv* '*cenruan 3*' elite cultivars (750 kg·ha^{-1}) (Liang et al. 2017). In plants, chloroplasts play an important role in maintaining life on Earth by providing carbohydrates, amino acids, lipids, and other metabolic substances (Daniell et al. 2021). Plant oil is one of the most important products of photosynthetic carbon assimilation. Fatty acid's biosynthesis occurred early in seed-filling stage and went on until seed maturing. Then, oil accumulated rapidly in seed at late stage of seed maturing (Cao et al. 2021). Previous studies show that acetyl-CoA carboxylase (ACCase) in plastids was a key enzyme regulating the rate of de novo fatty acid biosynthesis.

✉ Hang Ye
gfri_yehang@163.com

[1] Guangxi Key Laboratory of Special Non-Wood Forest Cultivation & Utilization; Guangxi Improved Variety and Cultivation Engineering Research Center of Oil-Tea Camellia, Chinese National Engineering Research Center for Oiltea Camelllia Tropical Area Experimental Station, Guangxi Forestry Research Institute, Nanning 530002, People's Republic of China

And the expression of the ACCase gene was directly correlated with change of lipid content (Modiri et al. 2018). Besides, the expression of oil biosynthesis-related transcription factors was influenced by the photosynthetic activity, such as WRINKLED1 (Hua et al. 2012). Therefore, research on oil biosynthesis and photosynthetic characteristic-related genes based on the whole chloroplast genome sequence of *C. osmantha cv* 'yidan' is of great significance for improving production. Moreover, the study of chloroplast genome genes provides a new idea for improving oil production in other oil plants.

At present, the chloroplast genome sequences of more than 20 plants in the genus *Camellia* have been published in NCBI, including species for ornamental purposes (Huang et al. 2013; Yang et al. 2013) and tea production and *C. oleifera*. The chloroplast (cp) genome is independent of the nuclear genome and exhibits maternal inheritance and semi-autonomous genetic characteristics (Guo et al. 2018). The structure of the cp genome in *Camellia* species is a typical four-segment, closed-loop structure, with a large single-copy (LSC) region, a small single-copy (SSC) region, and two inverted repeats (IRs) of roughly the same length (Zheng et al. 2019). Among these structural regions, the IRs are the most stable, and the LSC has a higher mutation rate than the SSC. The coding regions of genes have a slower evolution rate, which is suitable for the analysis of relationships at the family and higher levels, while the non-coding regions have a faster mutation rate (Chen et al. 2018), which is more suitable for analyzing relationships at lower levels such as genera and species (Clegg et al. 1994; Cui et al. 2019; Yang et al. 2019; Zeng et al. 2017). Thus, the characteristics of the maternal and highly conserved genes of the chloroplast genome provide favorable conditions for studying the phylogeny of plants.

Research on the chloroplast genome of *Camellia* plants is currently limited to the use of some chloroplast genes for phylogenetic analysis. Here, we describe the whole chloroplast genome sequence of *C. osmantha cv* 'yidan' and three other Camellia species using the next-generation Illumina genome analyzer platform. The three representative species have notable phenotypic differences (including pericarp thickness, fruit size, seed yield, and oil content) and are widely cultivated in southern China. This study aimed to provide more information for the classification of *C. osmantha cv* 'yidan' by clarifying and comparing the cp genome sequences and structural variations between *C. osmantha cv* 'yidan' and three closely related *Camellia* species.

2 Materials and methods

Sample preparation, sequencing, and chloroplast genome assembly –Fresh and healthy leaves of four Camellia species (*C. osmantha cv* 'yidan', *Camellia vietnamensis cv* 'hongguo', *Camellia oleifera cv* 'cenruan 3', and *Camellia semiserrata cv* 'hongyu 1') were sampled and used for complete cp genome sequencing. The four Camellia species were deposited in the Camellia oil Germplasm Resource (Latitude 22°55′51″, Longitude108°20′03″). A modified CTAB method was used to extract total genomic DNA from 50 mg of fresh leaves [58]. A 270- or 350-bp insertion library was constructed for each species, using TruSeq DNA sample preparation kits (San Diego, CA 92122 USA). DNA from the 4 species was indexed by tags and pooled for sequencing in Illumina PE (2 × 150 bp) at Kunming Institution of Botany, Chinese Academy of Sciences.

A total of 72 million raw reads were generated and made available in FASTQ format. The quality of the raw sequence reads was evaluated using the software package FastQC (Andrews 2010). The software Trimmomatic v0.36 was used for removal of adapter, contaminant, low-quality (Phred scores < 30), and short (< 36 bp) sequencing reads. The remaining high-quality sequencing reads were assembled de novo using the NOVOPlasty pipeline v2.7.2 with default parameters and based on a kmer size of 39 or 23 following the developer's suggestions, where the *psbA* gene of *C. oleifera cv* 'cenruan 3' was used as a seed input.

Chloroplast genomic annotation and sequence analyses – The assembled genomes of four species were originally annotated using PGA (Qu et al. 2019). The annotation results of codon positions and intron/exon boundaries were manually corrected by comparing with other known homologous genes (NC_023084.1) in the *Camellia* cp genome. The circular structures were mapped using the OGDRAW tool (Lohse et al. 2013). By aligning the IR/LSC and IR/SSC regions with homologous sequences from other *Camellia* species (NC_023084.1), their exact boundaries were determined.

Variation detection and evolutionary relationship analysis.

Repeat structures including palindromic, forward, complement, and reverse repeats were searched with BiBiServ software (https://bibiserv.cebitec.uni-bielefeld.de/reputer) with a repeat size of 15 bp and 90% or greater sequence identity. SSRs within the four cp genomes were detected using MISA software (https://webblast.ipk-gatersleben.de/misa/index.php). The following parameters were set in MISA: maximum length of sequence between two SSRs to register as compound SSR for 100 bp, with the parameters set at 10 for mononucleotides, 6 for dinucleotides, 5 for trinucleotides, and 5 for tetranucleotide, pentanucleotide, and hexanucleotide repeats.

We aligned the 114 *Camellia* and four other oil-producing species cp genome sequences using ClustalX. Unambiguously aligned DNA sequences were used for phylogenetic analyses, but ambiguously aligned regions were excluded. Maximum likelihood (ML) analyses were conducted using MEGA7. Bootstrap support (BS) values for individual

Table 1 The list of accession number of the chloroplast genome sequences used in this study

Taxon	GenBank accession number	Taxon	GenBank accession number
R.communis	NC_016736.1	*C.chrysanthoides*	MW543443.1
O.europaea	NC_013707.2	*C.achrysantha*	MW543442.1
R.communis	JF937588.1	*C.brevistyla*	MW256435.1
O.europaea	GU931818.1	*C.pubipetala*	MW186719.1
C.crapnelliana	KF753632.1	*C.perpetua*	MW186718.1
C.sinensis	KF562708.1	*C.sinensis var. sinensis cultivar Tieguanyin*	MW148820.1
C.taliensis voucher HKAS:S.X.Yang3157	KF156839.1	*C.sinensis isolate JM007 cultivar Bantianyao*	MW046255.1
C.yunnanensis voucher HKAS:S.X.Yang1090	KF156838.1	*C.fascicularis*	MW026668.1
C.pitardii voucher HKAS:S.X.Yang3148	KF156837.1	*C.meiocarpa*	MT956593.1
C.taliensis voucher HKAS:S.X.Yang3158	KF156836.1	*C.sinensis cultivar Tieluohan*	MT773377.1
C.impressinervis voucher HKAS:S.X.Yang1080	KF156835.1	*C.sinensis cultivar Shuijingui*	MT773376.1
C.danzaiensis voucher HKAS:S.X.Yang3147	KF156834.1	*C.sinensis cultivar Rougui*	MT773375.1
C.cuspidata voucher HKAS:S.X.Yang3159	KF156833.1	*C.grandibracteata*	NC_024659.1
C.sinensis	KC143082.1	*C.crapnelliana*	NC_024541.1
C.arabica	NC_008535.1	*C.yunnanensis voucherHKAS:S.X.Yang1090*	NC_022463.1
C.azalea	KY856741.1	*C.pitardii voucher HKAS:S.X.Yang3148*	NC_022462.1
C.luteoflora voucher CLUTE20161220	KY626042.1	*C.impressinervis voucherHKAS:S.X.Yang1080*	NC_022461.1
C.liberofilamenta voucher CLIBE20161220	KY626041.1	*C.danzaiensis voucherHKAS:S.X.Yang3147*	NC_022460.1
C.huana voucher CHUAN20161220	KY626040.1	*C.cuspidata voucher HKAS:S.X.Yang3159*	NC_022459.1
C.japonica	KU951523.1	*C.taliensis voucher HKAS:S.X.Yang3157*	NC_022264.1
C.sinensis var. sinensis	KJ806281.1	*C.sinensis*	NC_020019.1
C.sinensis var. pubilimba	KJ806280.1	*C.japonica strain Huaheling*	MW602996.1
C.sinensis var. dehungensis	KJ806279.1	*C.debaoensis*	MW543445.1
C.reticulata	KJ806278.1	*C.pubipetala*	MW543444.1
C.pubicosta	KJ806277.1	*C.nitidissima*	NC_039645.1
C.petelotii	KJ806276.1	*C.gymnogyna*	NC_039626.1
C.leptophylla	KJ806275.1	*C.ptilophylla*	NC_038198.1
C.grandibracteata	KJ806274.1	*C.granthamiana*	NC_038181.1
C.gymnogyna	MH394406.1	*C.chekiangoleosa*	NC_037472.1
C.gymnogyna	MH394405.1	*C.japonica strain S288C*	NC_036830.1
C.gymnogyna	MH394404.1	*C.azalea*	NC_035574.1
C.gymnogyna	MH394403.1	*C.reticulata*	NC_024663.1
C.nitidissima	MH382827.1	*C.pubicosta*	NC_024662.1
C.renshanxiangiae	MH253889.1	*C.petelotii*	NC_024661.1
C.sinensis	MH042531.1	*C.leptophylla*	NC_024660.1
C.ptilophylla	MG797642.1	*C.kissii*	NC_053915.1
C.granthamiana	MG782842.1	*C.fascicularis*	NC_053896.1
C.chekiangoleosa	MG431968.1	*C.yuhsienensis*	NC_053622.1
C.japonica strain S288C	MF850254.1	*C.gauchowensis*	NC_053541.1
C.oleifera	MF541730.2	*C.brevistyla*	NC_052752.1
C.sinensis sangmok	LC488797.1	*C.amplexicaulis*	NC_051559.1
C.grandibracteata	KJ806274.1	*C.rhytidophylla*	NC_050389.1
C.lungzhouensis	MN579509.2	*C.fraterna*	NC_050388.1
C.tachangensis cultivar Xingyi6	MN327576.1	*C.anlungensis voucher CANLU20191106*	NC_050354.1
C.sinensis cultivar Baiye 1	MN086819.1	*C.renshanxiangiae*	NC_041672.1
C.weiningensis voucher CwCPF1-201,901	MK820035.1	*C.sasanqua*	NC_041473.1
C.japonica isolate Jeju Island	MK353211.1	*C.sinensis var. assamica*	MH394407.1
C.japonica isolate Soyeonpyeongdo	MK353210.1	*C.sinensis cultivar Dahongpao*	MT773374.1

Table 1 (continued)

Taxon	GenBank accession number	Taxon	GenBank accession number
C.sasanqua	MH782189.1	*C.sinensis cultivar Baijiguan*	MT773373.1
C.sinensis	MH460639.1	*C.yuhsienensis*	MT665973.1
C.sinensis var. assamica	MH394410.1	*C.rhytidophylla*	MT663343.1
C.sinensis var. assamica	MH394409.1	*C.fraterna*	MT663342.1
C.sinensis var. assamica	MH394408.1	*C.chuongtsoensis*	MT663341.1
C.amplexicaulis	MT317095.1	*C.sinensis cultivar Wuyi Narcissus*	MT612435.1
C.anlungensis voucher CANLU20191106	MN756594.1	*C.gauchowensis*	MT449927.1
C.brevistyla	MN640791.1	*C.kissii*	MN635793.1

clades were calculated by running 1,000 bootstrap replicates of the data. ML Heuristic method searches were conducted with the nearest-neighbor-interchange (NNI). The genetic relationship of the four *Camellia* cp genomes together with 108 available *Camellia* (Table 1) and four other oil-producing species cp genome sequences (GenBank accession no. JF937588.1(*Ricinus communis* cultivar Hale), NC_016736.1(*Ricinus communis*), GU931818.1(*Olea europaea* cultivar Frantoio), and NC_013707.2) (*Olea europaea* cultivar Bianchera) were used to construct a maximum likelihood method (ML) tree by using MEGA 7 with default parameters (Tamura et al. 2011).

3 Results

The structure of the chloroplast genomes of four camellia species –The complete cp genomes of *C. semiserrata cv* ‘hongyu 1’ (GenBank accession no. OP953553), *C. vietnamensis cv* ‘hongguo’ (GenBank accession no. OP 953555), *C. osmantha cv* ‘yidan’ (GenBank accession no. OP936137), and *C. oleifera cv* ‘cenruan 3’ (GenBank accession no. OP953554) were sequenced using Illumina sequencing technology (Fig. 1). The cp genomes of the four species are composed of a circular DNA molecule ranging in size from 156,807 to 157,005 bp, with the typical quadripartite structure consisting of two inverted repeats (IRa and IRb) and LSC and SSC regions (Table 2).

The *C. semiserrata cv* ‘hongyu 1’, *C. osmantha cv* ‘yidan’, and *C. oleifera cv* ‘cenruan 3’ cp genomes each contain 134 genes (81 protein-coding genes, 39 transfer RNA (tRNA) genes, and 8 ribosomal RNA (rRNA) genes, as well as 6 genes with unknown functions). The *C. vietnamensis cv* ‘hongguo’ cp genome contains 136 genes (83 protein-coding genes, 39 tRNA genes, and 8 rRNA genes, as well as 6 genes with unknown functions), which includes two copies of the *rpl2* gene. By contrast, *rpl2* is not found in the other three species.

Among the 134 unique genes in *C. semiserrata cv* ‘hongyu 1’, *C. osmantha cv* ‘yidan’, and *C. oleifera cv* ‘cenruan 3’, 15 contain one intron (*petB*, *petD*, *atpF*, *ndhA*, *ndhB*, *rps12*, *rps16*, *rpl16*, *trnG-UCC*, *trnK-UUU*, *trnL-UAA*, *trnA-UGC*, *trnI-GAU*, *trnV-UAC*, and *rpoC1*), and 2 contain two introns (*clpP* and *ycf3*) (Table 3). Previous studies reported that ycf3 is necessary for the stable accumulation of the photosystem I complex (Boudreau et al. 1997; Naver et al. 2001; Guo et al. 2018). Among the 135 unique genes in *C. vietnamensis cv* ‘hongguo’, 16 contain one intron (*petB*, *petD*, *atpF*, *ndhA*, *ndhB*, *rps12*, *rps16*, *rpl2*, *rpl16*, *trnG-UCC*, *trnK-UUU*, *trnL-UAA*, *trnV-UAC*, *trnA-UGC* ,*trnI-GAU*, and *rpoC1*), and 2 contain two introns (*clpP* and *ycf3*). The gene maps of *C. osmantha cv* ‘yidan’, *C. semiserrata cv* ‘hongyu 1’, *C. oleifera cv* ‘cenruan 3’, and *C. vietnamensis cv* ‘hongguo’ are shown in Fig. 1.

Expansion and contraction of the border regions –The border regions and neighboring genes of the four *Camellia* cp genomes were compared to analyze the expansion and contraction of the connected regions (Fig. 2). The cp genomic structures, including gene type, gene order, and gene number, were conserved in *C. osmantha cv* ‘yidan’and *C. oleifera cv* ‘cenruan 3’, while the cp genomes of *C. vietnamensis cv* ‘hongguo’ exhibited visible differences at the IRb/SSC/IRa/borders. The IRb region expanded into the gene *ycf1* with 1042–1068 bp in the IRb regions (1068 bp for *C. osmantha cv* ‘yidan’ and *C. oleifera cv* ‘cenruan 3’, 1042 bp for *C. semiserrata cv* ‘hongyu 1’).

The IRa/SSC borders displayed large differences among the four cp genomes. The gene *ndhF* is located at the IRa/SSC or IRb/SSC junction, with 5–65 bp gaps between *ndhF* and the IR/SSC junction (5, 56, and 65 bp gaps in *C. semiserrata cv* ‘hongyu 1’, *C. osmantha cv* ‘yidan’, and *C. oleifera cv* ‘cenruan 3’, respectively). The *ndhF* and *ycf1* genes in *C. vietnamensis cv* ‘hongguo’ are reversed in the IRb/SSC/IRa boundary region compared with the cp genome sequences of the other three species. *ndhF* in the SSC region was 56 bp from the IRb/LSC junction in *C. vietnamensis cv*

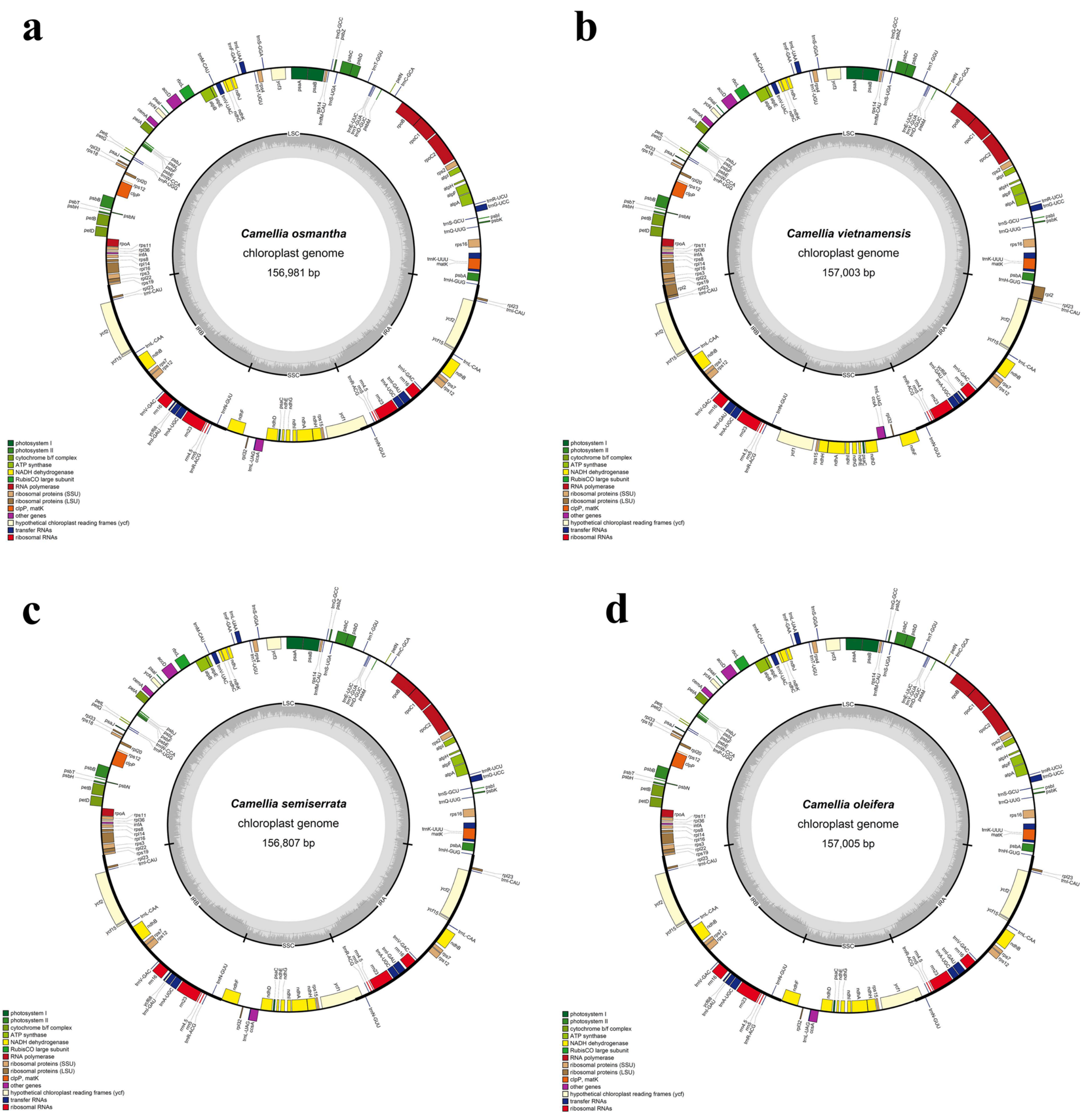


Fig. 1 Gene maps of the *C. osmantha*, *C. semiserrata*, *C. vietnamensis*, and *C. oleifera* cp genome. Genes drawn outside the circle are transcribed clockwise and those inside are transcribed counter clockwise. Genes belonging to different functional groups are color-coded. The inner dark gray represents the GC content of the chloroplast genome, and the light gray indicates the AT content. (Lohse et al. 2013)

'hongguo'. By contrast, the IRa/LSC and IRb/LSC boundary regions were relatively conserved in the four cp genomes. The gene *rpl2* formed another boundary by expanding into the IRa region in *C. vietnamensis cv* 'hongguo', leading to complete duplication of the gene within the IRs (Table 3).

Long-repeat and simple sequence repeat (SSR) analysis – We detected palindromic, forward, complementary, and reverse repeats in the four cp genomes. Overall, 50 repeat sequences were identified in all *Camellia* cp genomes, of which 23–24 palindromic repeats, 16–17 forward repeats, 7–9 reverse repeats, and 2–4 complementary repeats were separately found (Figure S1(A)). The lengths of palindromic repeats ranged from 19 to 79 bp, the forward repeats ranged in length from 19–42 bp, the reverse repeats ranged in length

Table 2 Summary of Camellia chloroplast genome features

	Camellia osmantha	*Camellia vietnamensis*	*Camellia semiserrata*	*Camellia oleifera*
Genome size (bp)	156,981	157,003	156,807	157,005
LSC size (bp)	86,647	86,656	86,449	86,632
SSC size (bp)	18,284	18,297	18,256	18,291
IRa size (bp)	26,025	26,025	26,051	26,041
IRb size (bp)	26,025	26,025	26,051	26,041
Number of genes	134	136	134	134

SSC (Small Single-Copy Region); IRs (Inverted Repeats Region); LSC (Large Single-Copy Region)

from 19–23 bp, and the complementary repeats ranged in length from 19–20 bp (Figure S1(B–E)).

In this study, we found 50, 51, 51, and 53 SSRs in the *C. semiserrata cv* 'hongyu 1*'*, *C. osmantha cv* 'yidan*'*, *C. vietnamensis cv* 'hongguo*'*, and *C. oleifera cv* 'cenruan 3*'* cp genomes, respectively (Fig. 3). These SSRs were mainly composed of adenine (A) or thymine (T) repeats and did not contain guanine (G) or cytosine (C) repeats. Moreover, the four cp genomes only contained mononucleotide repeats ranging from 10 to 17 bp.

Phylogenetic analysis –We generated a phylogenetic tree using the nucleotide sequences of the cp genomes of 112 *Camellia* species and other oilseed crops using the maximum likelihood method (Fig. 4), and *Coffea arabica* (NC_008535.1) was selected as an outgroup. *C. osmantha cv* 'yidan*'* is most closely related to *C. vietnamensis cv* 'hongguo*'* and *C. oleifera cv* 'cenruan 3*'*, which belong to the section *Oleifera* Chang.

4 Discussion

In this study, we sequenced the complete cp genomes of four *Camellia* species and annotated their sequences. Phylogenetic studies have shown that cp genome evolution includes nucleotide substitutions and structural changes (Feng et al. 2008; Haberle et al. 2008; Guo et al. 2018).

Some studies have shown that there are introns or gene deletions in the chloroplast genome (Downie et al. 1996; Downie et al. 1991; Graveley et al. 2001;Guisinger et al. 2010; Jansen et al. 2007; Ueda et al. 2007). Introns play an important role in the regulation of gene expression (Xu et al. 2017). They can increase gene expression levels in specific locations and at specific times (Niu et al. 2011; Le et al. 2003). The intron regulation mechanism has also been researched in other species (Callis et al. 1987; Emami et al. 2013). However, no studies have analyzed the association between intron loss and gene expression. The *chlB*, *chlL*, *chlN*, and *trnP-GGG* genes were missing in the four *Camellia* cp genomes but were found in several other angiosperm plastomes (Jansen et al. 2007; Green 2011; Mader et al. 2018). These four genes represent synapomorphies for flowering plants(Jansen et al. 2007). We found 15 genes that contained one intron and two genes that contained two introns (*ycf3* and *clpP*) in the *C. osmantha cv* 'yidan*'* cp genomes. The ycf3 protein is necessary to stabilize the complex of photosystem I with the light-harvesting complex I (Boudreau et al. 1997; Naver et al. 2001). We therefore speculate that intron gain in *ycf3* may alter the expression of genes encoding the photosystem I assembly protein. In the next study, we will focus on the photosynthesis-related genes in the four species. The *clpP* gene includes two introns. The intron gain in *clpP* may alter the regulation of genes encoding the clp protease proteolytic subunit. This phenomenon might be due to the increased evolutionary rates.

In addition, key genes related to lipid synthesis and photosynthesis are present in the chloroplast genome or located in chloroplast, such as carboxylase (*accD*) (Modiri et al. 2018), ω3-fatty acid desaturases(*FAD*) (Raboanatahiry et al. 2021), fatty acid exporter (*FAX1-1, FAX2, FAX4*) (xiao et al. 2021; Li et al. 2020), and phosphoenolpyruvate/phosphate translocator (*PPT*) genes (Tang et al. 2022). The *accD* gene encodes the heteroacetyl coenzyme A carboxylase (ACCase), a key enzyme involved in plant fatty acid biosynthesis (Nakkaew et al. 2008; Wicke et al. 2011; Kode et al. 2005; Zhang et al. 2016). Maliga (Maliga and Svab 2011) showed that *accD* in Nicotiana sylvestris was 1539 bp long. The *accD* sequence lengths were 1541, 1541, 1541, and 1532 bp in *C. oleifera cv* 'cenruan 3', *C. semiserrata cv* 'hongyu 1', *C. osmantha cv* 'yidan', and *C. vietnamensis cv* 'hongguo', respectively, suggesting that this gene has been conserved in plant cp genomes. Moreover, we observed no pseudogene formation of *accD* in the four Camellia cp genomes, consistent with the importance of fatty acid biosynthesis for these oil-producing plants. *Camelina sativa* ω3-fatty acid desaturases *CsaFAD7* and *CsaFAD8* were located in the chloroplast, which can modify the fatty acid composition of seed oil, which is useful for genetic engineering strategies (Raboanatahiry et al. 2021). *FAX1-1*, *FAX2*, and *FAX4* were both localized to the chloroplast membrane, which play critical roles in transporting plastid fatty acids for triacylglycerols (TAGs)

Table 3 List of genes in the three Camellia chloroplast genomes

Group of genes		Gene names	Number
Protein-coding genes	Large subunit of Rubisco	rbcL	1
	Photosystem1	psaA, psaB, psaC, ycf1, psaI	5
	PhotosystemII	psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ	15
	Cytochrome b/f complex	petA, petB*, petD*, petG, petL, petN	6
	ATP synthase	atpA, atpB, atpE, atpF(*), atpH, atpI	6
	NADH dehydrogenase	ndhA*, ndhB(2)*, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK	12
	Envelope membrane protein	cemA	1
	ATP-dependent protease subunit P	clpP**	1
Ribosomal proteins	Ribosomal small proteins	rps2, rps3, rps4, rps7(2), rps8, rps11, rps12(3)*, rps14, rps15, rps16*, rps18, rps19	15
	Ribosomal large proteins	rpl14, rpl16*, rpl20, rpl22, rpl23(2), rpl32, rpl33, rpl36	9
RNA genes	tRNA genes	trnA-UGC(2)*, trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnG-UCC*, trnG-GCC, trnH-GUG, trnI-CAU(2), trnI-GAU(2)*, trnK-UUU*, trnL-CAA(2), trnL-UAA*, trnL-UAG, trnM-CAU(2), trnN-GUU(2), trnP-UGG, trnV-GAC(2), trnQ-UUG, trnR-ACG(2), trnR-UCU, trnS-GGA, trnS-GCU, trnS-UGA, trnT-UGU, trnT-GGU , trnV-UAC*, trnV-GAC(2), trnW-CCA, trnY-GUA	39
	rRNA genes	rrn4.5(2), rrn5(2), rrn16(2), rrn23(2)	8
Transcription/ translation	Maturase	matK	1
	Subunit of acetyL-CoA carboxylase	Accd	1
	Functions unknown (conserved open reading frames)	ycf1, ycf2(2), ycf3**, ycf4, ycf15(2), ycf68	8
	c-type cytochrome synthesis	ccsA	1
	DNA-dependent RNA polymerase	rpoA, rpoB, rpoC1*, rpoC2	4
	Translational initiation factor	infA	1
Total			134

rpl2: 2 copies in *C. vietnamensis* and 0 in the other three species

[*]genes containing one intron; **genes containing two introns; (2) genes present in two copies; (3) genes present in three copies

biosynthesis during seed embryo development (Li et al. 2020). *BnaFAX1-1* may simultaneously improve seed oil content, oil quality, and biological yield in *B. napus* (xiao et al. 2021). *BnaPPT1* plays an important role in leaf membrane lipid synthesis and chloroplast development, thus affecting photosynthesis (Tang et al. 2022). Therefore, the study of lipid metabolism-related genes in the chloroplast genome provides a new approach for future molecular breeding in *camellia* oil.

Previous studies showed that *C. oleifera cv* 'cenruan 3' is more adapted to low light conditions compared to the other Camellia species (Ma et al. 2012a, b). And, the light saturation point of *C. osmantha cv* 'yidan' is 499.7 μmol · m^{-2} s^{-1}, and this species is more adapted to high light conditions. So, the light energy utilization of *C. osmantha* is maybe higher. Differences in plant photosystems maybe used to improve the efficiency of light absorption and transformation and further increase plant yield (Zhang et al. 2011). As the center of photosynthesis, the chloroplast genome is of great significance for revealing the mechanism and metabolic regulation of plant photosynthesis (Fang et al. 2010; Huang et al. 2013). Seed or silique wall photosynthesis contributed to the increased seed weight and oil content (Hu et al. 2018; Liu et al. 2012). The *rpoA* and *rpoC2* genes encode the alpha and beta subunits of plastid RNA polymerase (PEP), respectively, which is responsible for the transcription of most photosynthetic proteins. We speculate that *rpoA* and *rpoC2* genes in the chloroplast genome play a key role in the photosynthesis of *C. osmantha*.

Besides, it has been shown that when using chloroplast gene fragments for species low-order unit delineation, applicable highly variable regions should first be screened in the whole chloroplast genome (Dong et al. 2012). Chloroplast molecular markers in hypervariable region analysis can explain the intraspecific divergences in the species (Lin et al. 2022; Li et al. 2022; Xiong et al. 2022). Moreover, chloroplast genomes can develop a high-resolution molecular marker for tracking population genetic diversity (Song et al. 2020). In *C. vietnamensis cv* 'hongguo', *rpl2* is present and has not been found in the other three species. The gene

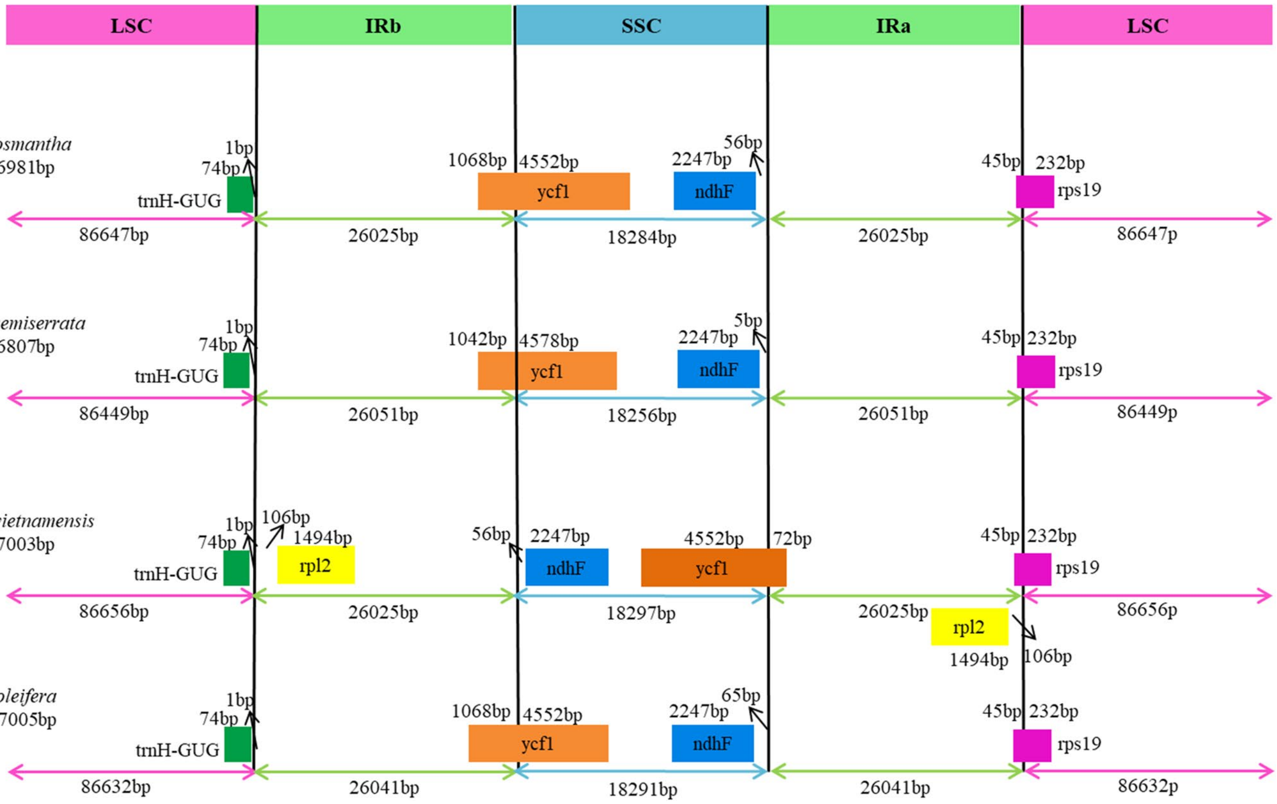
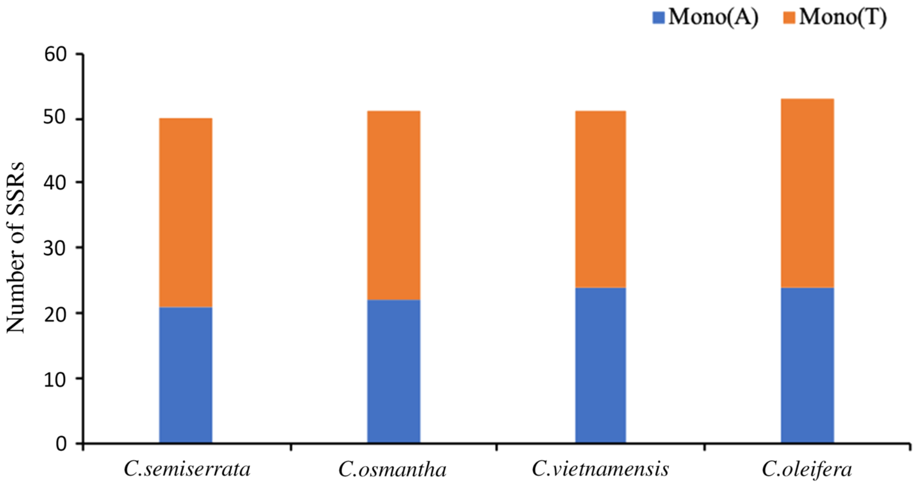


Fig. 2 Comparison of the SSC, IRs, and LSC border regions among the four *Camellia* cp genomes. *Note*: SSC(Small Single-Copy Region); IRs(Inverted Repeats Region); LSC(Large Single-Copy Region)

Fig. 3 Number of SSR motifs in different Camellia cp genomes



encodes a ribosomal protein L2, which full-length sequence is 1494 bp with a 671 bp intron. The *rpl2* is found in other plants of the genus Camellia, so the development of molecular markers using the rpl2 gene could be used to distinguish thee four species, but whether it can be used to differentiate them from other Camellia spp. and requires further research.

Phylogenetic relationships among four *Camellia* species revealed that *C. osmantha cv* 'yidan*'* is more closely related to *C. vietnamensis cv* 'hongguo*'* and *C. oleifera cv* 'cenruan 3*'* than to *C. semiserrata cv* 'hongyu 1*'*, other *Camellia* species, and other oil crops. The results of this study provide an assembly of a whole chloroplast genome of *C. osmantha cv 'yidan'*, which may be useful for future breeding and further biological discoveries. It will provide a theoretical basis for the improvement of Camellia oil yield and the determination of phylogenetic status.

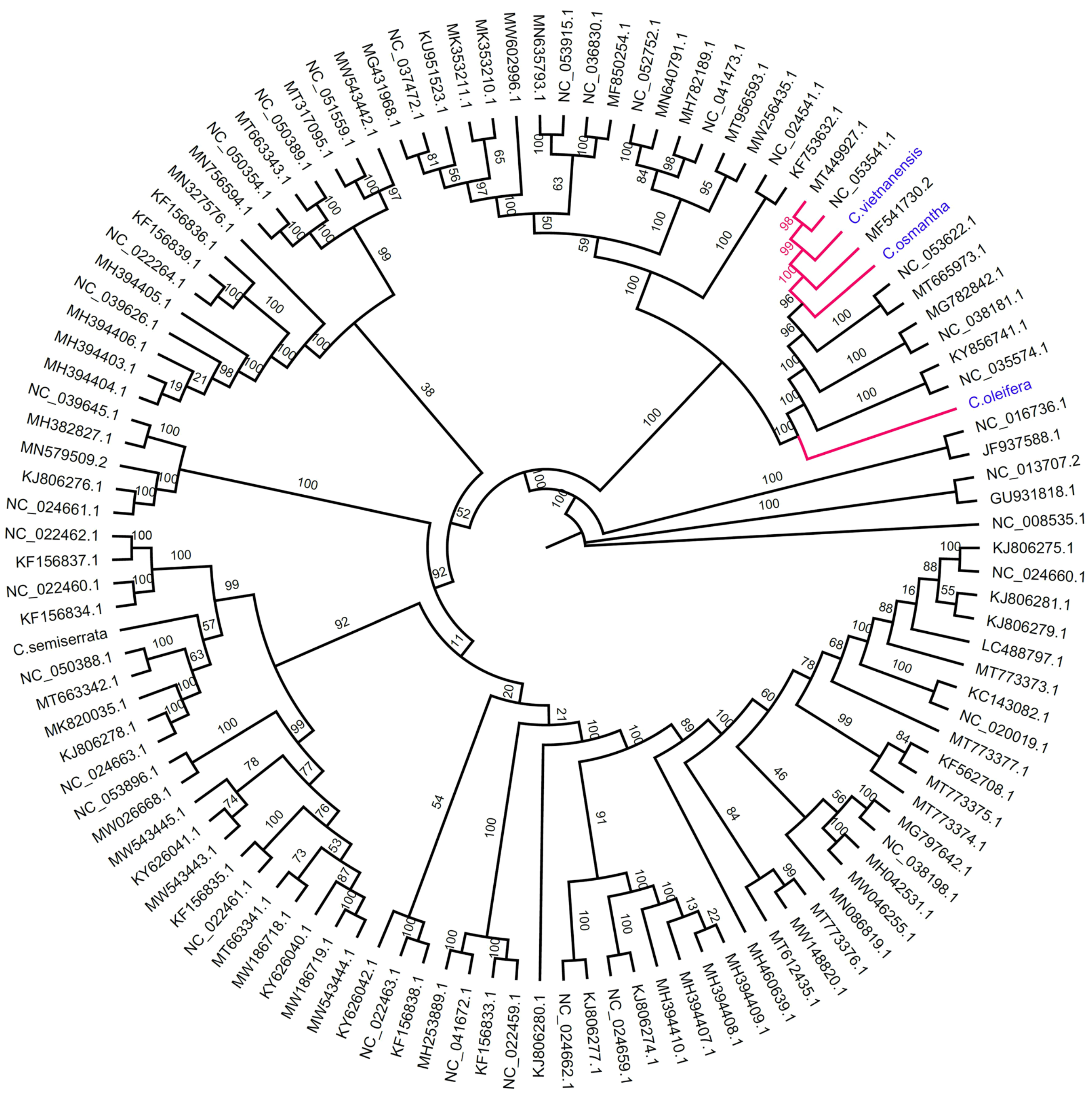


Fig. 4 Phylogenetic tree of *Camellia* and other related oilseed species by using the maximum likelihood method

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40415-023-00960-2.

Acknowledgements Thanks to Hong Yang for the analysis, assembling and annotation of sequencing sequences raw reads in the Kunming Institute of Botany, CAS. The sequencing of the four Camellia species involved in the article was performed on the Germplasm Bank of Wild Species.

Author contributions The research structure was designed by JM and HY; BH prepared the sample and performed the experiments, analyzed the data and wrote the paper; GC and JM made revisions to the final manuscript. The final manuscript was read and corrected by all authors.

Funding This work was supported by Natural Science Foundation of Guangxi Province under grant numbe 2021GXNSFBA196085 and Specific Research Project of Guangxi for Research Bases and Talents under grant number AD20325005, (National Natural Science Foundation of China, under Grant #31960322 (funder: National Natural Science Foundation Committee of China).

Data availability All the data involved in this article are true and reliable. The specimens of four Camellia species plants are deposited in the Camellia oleifera Germplasm Resource (http://www.gxlky.com.cn/, Bingqing Hao, bingqinghao_2021@caf.cn). The DNA samples are stored in the laboratory of Guangxi Academy of Forestry(http://www.gxlky.com.cn/, Bingqing Hao, bingqinghao_2021@caf.cn). The genome sequence data that support the findings of this study are openly available in the NCBI. *C. semiserrata cv* 'hongyu 1'(GenBank accession no. OP953553), *C. vietnamensis cv* 'hongguo' (GenBank accession no. OP 953555), *C. osmantha cv* 'yidan'(GenBank accession no. OP936137), and *C. oleifera cv* 'cenruan 3'(GenBank accession no. OP953554).

Declarations

Conflict of interest The authors are really grateful to the opened raw genome data from public database. The authors report no conflicts of interest and are responsible for the content and writing of the paper.

Ethical approval and consent to participate Our study does not involve ethics approval and consent to participate.

Consent to publish All authors read and approved the final manuscript.

References

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge

Boudreau E, Takahashi Y, Lemieux C et al (1997) The chloroplast ycf3 and ycf4 open reading frames of of Chlamydomonas reinhardtii are required for the accumulation of the photosystem I complex. EMBO J 16:6095–6104

Callis J, Fromm M, Walbot V (1987) Introns increase gene expression in cultured maize cells. Genes Dev 1:1183–1200

Cao J, He L, Nwafor CC, Qin L et al (2021) Ultrastructural studies of seed coat and cotyledon during rapeseed maturation. J Integr Agric 20:1239–1249

Clegg MT, Gaut BS, Learn GH et al (1994) Rates and patterns of chloroplast DNA evolution. Proc Natl Acad Sci USA 91:6795–6801

Cui Y, Zhou J, Chen X et al (2019) Complete chloroplast genome and comparative analysis of three Lycium (Solanaceae) species with medicinal and edible properties. Gene Reports 17:100464

Chen X, Zhou J, Cui Y, Wang Y, Duan B and Yao H (2018) Identification of Ligularia herbs using the complete chloroplast genome as a super-barcode. Front Pharmacol 9:695

Daniell H, Jin S, Zhu XG et al (2021) Green giant-a tiny chloroplast genome with mighty power to produce high-value proteins: history and phylogeny. Plant Biotechnol J 19:430–447

Dong W, Liu J, Yu J et al (2012) Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. PLoS ONE 7:e35071

Downie SR, Llanas E, Katz-Downie DS (1996) Multiple independent losses of the rpoC1 intron in angiosperm chloroplast DNA's. Syst Bot 21:135–151

Downie SR, Olmstead RG, Zurawski G et al (1991) Six independent losses of the chloroplast DNA rpl2 intron indicotyledons: molecular and phylogenetic implications. Evolution 45:1245–1259

Emami S, Arumainayagam D, Korf I, Rose AB (2013) The effects of a stimulating intron on the expression of heterologous genes in Arabidopsis thaliana. Plant Biotechnol J 11:555–563

Fang W, Yang JB, Yang SX et al (2010) Phylogeny of Camellia sects. Longipedicellata, Chrysantha and Longissima (Theaceae) based on sequence data of four chloroplast DNA loci. Acta Bot Yunnanica 32:1–13

Feng Y, Cui L, Depamphilis CW et al (2008) Gene rearrangement analysis and ancestral order inference from chloroplast genomes with inverted repeat. BMC Genomics 9:35

Graveley BR (2001) Alternative splicing: Increasing diversity in the proteomic world. Trends Genet 17:100–107

Green BR (2011) Chloroplast genomes of photosynthetic eukaryotes. Plant J 66:34–44

Guisinger MM, Chumley TW, Kuehl JV, Boore JL, Jansen RK (2010) Implications of the plastid genome sequence of Typha (Typhaceae, Poales) for understanding genome evolution in Poaceae. J Mol Evol 70:149–166

Guo S, Guo L, Zhao W et al (2018) Complete chloroplast genome sequence and phylogenetic analysis of *Paeonia ostii*. Molecules 23:246

Haberle RC, Fourcade HM, Boore JL, Jansen RK (2008) Extensive rearrangements in the chloroplast genome of Trachelium caeruleum are associated with repeats and tRNA genes. J Mol Evol 66:350–361

Hu Y, Zhang Y, Yu W et al (2018) Novel insights into the influence of seed sarcotesta photosynthesis on accumulation of seed dry matter and oil content in *Torreya grandis cv*. "Merrillii." Front Plant Sci 8:2179

Hua W, Li RJ, Zhan GM et al (2012) Maternal control of seed oil content in *Brassica napus* the role of silique wall photosynthesis. Plant J 69:432–444

Huang H, Tong Y, Zhang QJ, Gao LZ (2013) Genome size variation among and within Camellia species by using flow cytometric analysis. PLoS ONE 8:e64981

Huang H, Shi C, Liu Y, Mao SY, Gao LZ (2014) Thirteen Camellia chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. BMC Evol Biol 14:151. https://doi.org/10.1186/1471-2148-14-151

Jansen RK, Cai Z, Raubeson LA et al (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci USA 104:19369–19374

Kode V, Mudd EA, Iamtham S, Day A (2005) The tobacco plastid accD gene is essential and is required for leaf development. Plant J 44:237–244

Le H, Nott A, Moore MJ (2003) How introns influence and enhance eukaryotic gene expression. Trends Biochem Sci 28:215–220

Li J, Yang M, Li Y et al (2022) Chloroplast genomes of two Pueraria DC. species: sequencing, comparative analysis and molecular marker development. FEBS Open Bio 12:349–361

Li N, Meng H, Li S et al (2020) Two plastid fatty acid exporters contribute to seed oil accumulation in *Arabidopsis*. Plant Physiol 182:1910–1919

Liang HY, Hao BQ, Chen GC et al (2017) Camellia as an oil-seed crop. Hort Science 52:488–497

Lin S, Liu J, He X et al (2022) Comprehensive comparative analysis and development of molecular markers for Dianthus species based on complete chloroplast genome sequences. Int J Mol Sci 23:12567

Liu J, Hua W, Yang HL et al (2012) The *BnGRF2* gene (GRF2-like gene from *Brassica napus*) enhances seed oil production through regulating cell number and plant photosynthesis. J Exp Bot 63:3727–3740

Liu K, Zhou ZD, Wang DX et al (2013) Flooding tolerance of five camellia species. Guangxi for Res Sci 42:329–332

Lohse M, DrechseL O, Kahlau S et al (2013) Organellar genome DRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Res 41:W575–W581

Ma JL (2020) YiDan. Guangxi Forestry Research Institute, Guangxi, 2020–03–02

Ma JL, Ye H, Ye CX (2012a) A new species of Camellia sect. Paracamellia Guangxi Plant 32:753–755

Ma JL, Zhang RQ, Ye H et al (2012b) Photosynthetic characteristics among different Camellia species. Nonwood Forest Res 30:73–76

Ma JL, Zhang RQ, Ye H, He XY (2013) Semi-lethal temperature and cold tolerance & heat tolerance in *Camellia osmantha*. Nonwood Forest Res 31:150–152

Mader M, Pakull B, Blanc-Jolivet C et al (2018) Complete chloroplast genome sequences of four Meliaceae species and comparative analyses. Int J Mol Sci 19:701

Maliga P, Svab Z (2011) Engineering the plastid genome of *Nicotiana sylvestris*, a diploid model species for plastid genetics. In plant chromosome engineering. Humana Press, Totowa, pp 37–50

Modiri S, Zahiri HS, Vali H, Noghabi KA (2018) Evaluation of transcription profile of acetyl-CoA carboxylase (*ACCase*) and acyl-ACP synthetase (*AAS*) to reveal their roles in induced lipid accumulation of Synechococcus sp. *HS01*. Renew Energy 129:347–356

Nakkaew A, Chotigeat W, Eksomtramage T, Phongdara A (2008) Cloning and expression of a plastid-encoded subunit, beta-carboxyltransferase gene (accD) and a nuclear-encoded subunit, biotin carboxylase of acetyl-CoA carboxylase from oil palm (Elaeis guineensis Jacq). Plant Sci 175:497–504

Naver H, Boudreau E, Rochaix JD (2001) Functional studies of *Ycf3*: its role in assembly of photosystem I and interactions with some of its subunits. Plant Cell 13:2731–2745

Niu DK, Yang YF (2011) Why eukaryotic cells use introns to enhance gene expression: splicing reduces transcription-associated mutagenesis by inhibiting topoisomerase I cutting activity. Biol Direct 6:24

Qu XJ, Moore MJ, Li DZ, Yi TS (2019) PGA: a software package for rapid, accurate, and flexible batch annotation of plastomes. Plant Methods 15:50

Raboanatahiry N, Yin Y, Chen K et al (2021) In silico analysis of fatty acid desaturases structures in *Camelina sativa*, and functional evaluation of *Csafad7* and *Csafad8* on seed oil formation and seed morphology. Int J Mol Sci 22:10857

Robards K, Prenzler P, Ryan D, Zhong H (2009) Camellia oil and tea oil. In: Moreau R, Kamal-Eldin (eds) Gourmet and health promoting specialty oils. AOCS Press, Urbana, IL, pp 313–343

Song H, Liu F, Li Z et al (2020) Development of a high-resolution molecular marker for tracking Phaeocystis globosa genetic diversity through comparative analysis of chloroplast genomes. Harmful Algae 99:101911

Tamura K, Peterson D, Peterson N et al (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–2739

Tang S, Peng F, Tang Q et al (2022) *BnaPPT1* is essential for chloroplast development and seed oil accumulation in *Brassica napus*. J Adv Res 42:29–40

Ueda M, Fujimoto M, Arimura SI (2007) Loss of the *rpl32* gene from the chloroplast genome and subsequent acquisition of a preexisting transit peptide within the nuclear gene in Populus. Gene 402:51–56

Vijayan K, Chung MC, Tsou CH (2012) Dispersion of rDNA loci and its implications on intragenomic variability and phylogenetic studies in Camellia. Sci Hort 137:59–68

Wang DX, Ye H, Ma JL, Zhou ZD (2014) Evaluation and selection of *Camellia osmantha* germplasm resources. Nonwood Forest Res 32:159–162

Wicke S, Schneeweiss GM, Depamphilis CW et al (2011) The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol 76:273–297

Xiao Z, Tang F, Zhang L et al (2021) The Brassica napus fatty acid exporter *FAX1-1* contributes to biological yield, seed oil content, and oil quality. Biotechnol Biofuels 14:190

Xiong Y, Xiong Y, Shu X et al (2022) Molecular phylogeography and intraspecific divergences in siberian wildrye (*Elymus sibiricus* L.) wild populations in China, inferred from chloroplast dna sequence and cpssr markers. Front Plant Sci 13:862759

Xu J, Chu Y, Liao B et al (2017) Panax ginseng genome examination for ginsenoside biosynthesis. Gigascience 6:1–15

Yang JB, Yang SX, Li HT, Yang J, Li DZ (2013) Comparative chloroplast genomes of Camellia species. PLoS ONE 8:e73053. https://doi.org/10.1371/journal.pone.0073053

Yang Z, Huang Y, An W et al (2019) Sequencing and structural analysis of the complete chloroplast genome of the medicinal plant *Lycium chinense* Mill. Plants 8:87

Zeng S, Zhou T, Han K et al (2017) The complete chloroplast genome sequences of six Rehmannia species. Genes 8:103

Zhang JM, Liu J, Sun HL et al (2011) Nuclear and chloroplast SSR markers in *Paeonia delavayi* (Paeoniaceae) and cross-species amplification in P. ludlowii. Am J Bot 98:346–348

Zhang Y, Du L, Liu A et al (2016) The complete chloroplast genome sequences of five Epimedium species: lights into phylogenetic and taxonomic analyses. Front Plant Sci 7:306

Zheng X, Ren C, Huang S et al (2019) Structure and features of the complete chloroplast genome of *Melastoma dodecandrum*. Physiol Mol Biol Plants 25:1043–1054

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