INTERNATIONAL JOURNAL OF AGRICULTURE & BIOLOGY
ISSN Print: 1560–8530; ISSN Online: 1814–9596
20–0077/2020/24–4–743–750
DOI: 10.17957/IJAB/15.1495
http://www.fspublishers.org

Full Length Article

Complete Chloroplast Genome Sequences and Phylogenetic Analysis of Three *Camellia oleifera* Cultivars

Lingli Wu[1,2], Jian'an Li[1,2], YiYang Gu[1,2], Fanhang Zhang[1,2], Ze Li[1,2*] and Xiaofeng Tan[1,2*]

[1]*Key Laboratory of Cultivation and Protection for Non-Wood Forest Trees of Ministry of Education and the Key Laboratory of Non-Wood Forest Products of Forestry Ministry, Central South University of Forestry and Technology, Changsha, 410004, China*

[2]*Engineering Technology Research Center of Southern Hilly and Mountainous Ecological Non-Wood Forest Industry of Hunan Province, Central South University of Forestry and Technology, Changsha, 410004, China*

[*]For correspondence: lize1853@163.com; tanxiaofengcn@126.com

Received 15 January 2020; Accepted 09 April 2020; Published 16 August 2020

Abstract

Camellia oleifera is an economically important edible oil tree that is cultivated over an area of about 4.3 million hectares in southern China. To elucidate the molecular and evolutionary characteristics of *C. oleifera* chloroplasts, we sequenced and characterized its genome using Illumina high-throughput sequencing (HiSeq) combined with third-generation PacBio sequencing technology. The chloroplast genomes ranged from 156,965 to 156,975 bp, containing 133 genes including 8 rRNAs, 37 tRNAs and 88 protein-coding genes, and had a typical quadripartite structure with conserved genome arrangement. Approximately 50 simple sequence repeats (SSRs) and 37 repeat sequences were identified in the *C. oleifera* chloroplast genome. A comparison with seven other *Camellia* complete chloroplast genomes detected only slight differences in the coding regions of some genes, including *psbN* and *ycf1*; highly divergent regions were observed in the intergenic spacers (IGS) of the 10 *Camellia* chloroplast genomes. Phylogenetic analysis suggested a sister relationship among three *C. oleifera* cultivars ('Huashuo', 'Huajin' and 'Huaxin') and *C. oleifera*, whereas *C. oleifera* from Hainan was identified as sister to *C. azalea*. Based on our evaluation of the chloroplast genomic resources of three *C. oleifera* cultivars, we assembled the complete chloroplast genomes; the development of abundant *C. oleifera* genetic resources will contribute to further studies of genetic engineering, phylogeny, and evolution among family *Theaceae* species. © 2020 Friends Science Publishers

Keywords: *Camellia oleifera*; Chloroplast genome; Phylogenetic relationship; Repeat analysis

Introduction

Camellia oleifera L. (Abel) is a woody oil tree of the genus *Camellia* in the family Theaceae. *C. oleifera* seeds are harvested for the extraction of an edible tea oil that has high nutritional value and healthful properties including blood cholesterol reduction and the prevention of hypertension and arteriosclerosis (Feás *et al*. 2013; Zeng *et al*. 2015; Qu *et al*. 2019). For the reasons that its chemical composition and unsaturated fatty acid contents are similar to those of olive oil, *C. oleifera* oil is known as "eastern olive oil" (Gao *et al*. 2015; Li *et al*. 2016). *C. oleifera* seed meal can be used to extract saponin for feed production, and the shells can be used to produce potassium carbonate or to cultivate edible and medicinal fungi (Hu *et al*. 2012; Zhu *et al*. 2018). *C. oleifera* is the most valued oil-producing plant in China (Tan *et al*. 2018). In recent years, *C. oleifera* has been planted over large areas in hilly regions of southern China with red soil (Wang *et al*. 2019).

C. Oleifera cultivars with desirable traits have been planted on a large scale by farmers. High-yield cultivars *C. oleifera* 'Huashuo', 'Huajin' and 'Huaxin' were bred from *C. oleifera* in 2009. *C. oleifera* 'Huashuo' has large fruit, high yield, strong resistance, and late maturity (Tan *et al*. 2011) (Fig. 1A); *C. oleifera* 'Huajin' has rapid growth, large green leaves, high yield, precocity, and pear-shaped fruit (Yuan 2012) (Fig. 1B); and *C. oleifera* 'Huaxin' has high and stable yield, strong resistance, precocity, and red fruit (Tan *et al*. 2012) (Fig. 1C). These three *C. oleifera* cultivars have been cultivated widely in the hilly red soil region of Hunan Province in recent years (Wu *et al*. 2020). However, the genetic backgrounds of these cultivars are poorly known, and genetic resources are scarce.

Chloroplasts are key organelles that act as the plant metabolic center; they contain the complete enzymatic machinery for plant growth and development, with carbon fixation and oxygen release. The chloroplast genome, one of three DNA genomes in the plant body, has a highly conserved circular DNA arrangement that encodes many key proteins related to photosynthesis (Bobik and Burch-

To cite this paper: Wu L, J Li, YY Gu, F Zhang, Z Li, X Tan (2020). Complete chloroplast genome sequences and phylogenetic analysis of three *Camellia oleifera* cultivars. *Intl J Agric Biol* 24:743–750

Smith 2015; Zhang *et al*. 2017; Liu *et al*. 2018). Since publication of the first chloroplast genome from *Marchantia polymorpha* (Kazuhiko *et al*. 1984; Wang *et al*. 2016), over 2,500 chloroplast genomes have been sequenced (http://www.ncbi.nlm.nih.gov/genomes/), providing insights into plant diversity, and evolution, and have been applied in DNA barcoding and genetic engineering of biomedical products (Kang *et al*. 2017; Song *et al*. 2017). Most chloroplast genomes range from 115 to 165 kb and have a quadripartite organization, including a large single-copy (LSC) region, a small single-copy (SSC) region and a pair of inverted repeats (IRs) (Li *et al*. 2017; Xu *et al*. 2017). Chloroplast genomes do not undergo recombination; they exhibit maternal inheritance and greater conservation than observed in nuclear and mitochondrial genomes (Palmer *et al*. 1988; Wu *et al*. 2010). *C. oleifera* is a widely distributed self-incompatible plant with extremely complex cross-pollination characteristics and intraspecific variation. The *C. oleifera* genome has not been sequenced, and most *C. oleifera* cultivars are polyploid, with complex genetic backgrounds and evolutionary processes. Moreover, the chloroplast genome sequences of the three *C. oleifera* cultivars have not yet been elucidated; therefore, it is important to clarify the phylogenetic and evolutionary relationships among different *Camellia* species to improve and expand the range of existing cultivars.

In this study, we assembled the complete chloroplast genome sequences of three important *C. oleifera* cultivars ('Huashuo', 'Huaxin' and 'Huajin') and characterized their genomes using Illumina high-throughput sequencing (HiSeq) technology. Genome maps of the obtained sequencing data were mapped using bioinformatics analysis to reveal the photosynthesis mechanisms and phylogenetic relationships of these cultivars relative to other *C. oleifera* cultivars. We also performed comparative analyses using known chloroplast genomes to improve our understanding of the *C. oleifera* chloroplast genome. The objective of this study was to investigate plant molecular markers, species relationships, and the structure and origin of chloroplast DNA, and to further explore the evolution of *Camellia* species using molecular methods.

Materials and Methods

Plant materials and DNA sequencing

Fresh leaves of the three *C. oleifera* cultivars were collected from 8-year-old trees growing at the Central South University of Forestry Science and Technology (112°40'E, 28°29'N, Wang Cheng, Hunan, China). Approximately 5 g of fresh leaves were harvested for chloroplast DNA isolation using an improved extraction method (McPherson *et al.* 2013). After DNA isolation, 1 μg purified DNA was fragmented to construct short-insert libraries (insert size, 430 bp) according to the manufacturer's instructions (Illumina) and then sequenced using the Illumina HiSeq4000 system (Borgstrom *et al*. 2011; Shanghai Biozeron Biotechnology). High-molecular-weight DNA was purified and used to prepare the PacBio library and for BluePippin size selection and then sequenced using a Sequel sequencer. Additional sequencing was performed by Nextomics (Wuhan, China) using the PacBio RS II platform.

Genome assembly

Before assembly, the Illumina raw reads were filtered to remove reads with adaptors, low-quality reads (Q<20), reads containing ≥10% N characters, and duplicate sequences. We assembled the genome framework using both Illumina and PacBio data with SPAdes v. 3.10.1 (Antipov*et al*. 2016). Next, we verified the assembly and circular character of the chloroplast genomes, filling any gaps that occurred.

Genome annotation

We annotated the chloroplast genes using the DOGMA online tool (Wyman *et al.* 2004). A whole-chloroplast genome blast search was performed using the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Minoru *et al*. 2014), Clusters of Orthologous Groups (COG) (Tatusov *et al.* 2003), Non-Redundant (NR) Protein, Swiss-Prot (Magrane and Consortium 2011) and Gene Ontology (GO) (Ashburner *et al.* 2000) databases. A circular chloroplast genome map was drawn using Organellar Genome DRAW v. 1.2 (Lohse *et al*. 2007).

Chloroplast genome sequence analysis

The programs Mauve (Ravi *et al*. 2006) and mVISTA programs were used to identify similarities among different chloroplast genomes (Mayor *et al*. 2000). The REPuter program was used to identify and locate forward (direct) repeats, reverse sequences, complementary sequences, and palindromic sequences with lengths of at least 22 bp and sequence identity ≥90% (Kurtz *et al.* 2001). SSR distributions were predicted using the MISA microsatellite search tool (Beier *et al*. 2017). IR expansion/contraction regions were compared among *C. oleifera* 'Huashuo', *C. oleifera* 'Huajin', *C. oleifera* 'Huaxin', *N. tabacum*, *C. sinensis*, *C. petelotii*, *C. pitardii* and *C. oleifera*.

Phylogenetic analysis

The maximum likelihood (ML) analysis was performed using RAxML v. 7.2.6 with the default parameters (Stamatakis 2006). The maximum parsimony (MP) analyses were performed using PAUP 4.0.

Results

General features of *C. oleifera* chloroplast DNA

The *C. oleifera* 'Huashuo', 'Huajin' and 'Huaxin'

Fig. 1: The three *Camellia oleifera* cultivars, fruiting in September. **A**: 'Huashuo'; **B**: 'Huajin'; **C**: 'Huaxin'

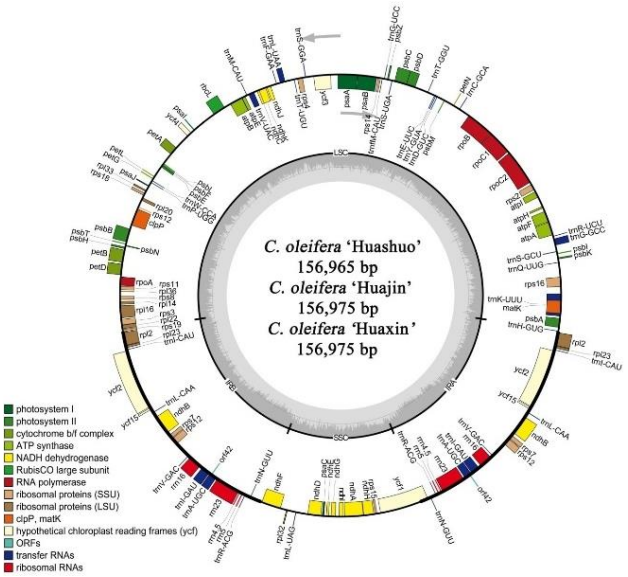

Fig. 2: Gene map of the three *C. oleifera* cultivar chloroplast genomes determined using the PacBio RS II platform. Thick lines indicate inverted repeats (IRa and IRb), which separate the genome into large single-copy (LSC) and small single-copy (SSC) regions. Genes on the outside of the map are transcribed in a clock wise direction; those inside the map are transcribed in a counter clock wise direction

chloroplast genomes were 156,965, 156,975 and 156,975 bp in length, respectively. These genomes were similar to those of most angiosperms, with an LSC region of 86,650 bp in 'Huashuo' and 86,660 bp in both 'Huajin' and 'Huaxin' and an SSC region of 18,409 bp in 'Huashuo' and 18,406 bp in both 'Huajin' and 'Huaxin' separated by a pair of IRs of 51,906 bp in 'Huashuo' and 51,908 bp in both 'Huajin' and 'Huaxin' (Fig. 2 and Table 1). The genomes had guanine–cytosine (GC) contents of 37.29%; however, the GC content was higher in the rRNA region (55.41% in all three cultivars) than in the overall genome (Table 1).

Gene content, orientation, and order were similar among the three *C. oleifera* cultivars. A total of 133 genes consisting of 88 protein-coding genes, 37 tRNA genes and 8 rRNA genes were identified from each genome (Table 1 and 2). A total of 20 genes (8 tRNA, 4 rRNA, and 8 protein-coding genes) were duplicated in the IR regions of each genome (Fig. 2). 'Huashuo' had 16 genes with introns, whereas both 'Huajin' and 'Huaxin' had only 15 genes with introns. GC protein-coding genes were 77,866 bp in length in 'Huashuo' and 76,648 bp in 'Huajin' and 'Huaxin'. Therefore, the final chloroplast genome sequences of the three *C. oleifera* were obtained (Submitted to the NCBI database).

Table 1: Characteristics of the plastome genomes of three *Camellia oleifera* cultivars. cp, chloroplast. LSC, large single copy; SSC, small single copy; IR, inverted repeat; IGS, intergenic spacer. GC, guanine–cytosine content

Sequence region	Length (bp)/Percent (%)		
	C. oleifera 'Huashuo'	*C. oleifera* 'Huaxin'	*C. oleifera* 'Huajin'
Total cp genome	156,965	156,975	156,975
LSC region	86,650	86,661	86,661
SSC region	18,409	18,406	18,406
IR region	51,906	51,908	51,908
Coding regions	79,500	79,504	79,504
Introns	110,882	110,879	110,879
rRNA	9,046	9,046	9,046
tRNA	2,802	2,802	2,802
IGS	77,357	77,364	77,364
GC content	Length (bp)/Percent (%)		
Overall GC size	58,532/37.29	58,535/37.29	58,537/37.29
Overall A size	48,745/31.05	48,749/31.06	48,752/31.06
Overall T size	49,688/31.66	49,691/31.66	49,686/31.65
Overall G size	28,677/18.27	28,678/18.27	28,681/18.27
Overall C size	29,855/19.02	29,857/19.02	29,856/19.02
GC content in protein-coding regions	77,866 (40.90)	76,648 (40.26)	76,648 (40.26)
GC content in introns	38.97	38.97	38.97
GC content in rRNA	5,012/55.41	5,012/55.41	5,012/55.41
GC content in tRNA	1,482/52.89	1,482/52.89	1,482/52.89
GC content in IGS	28,638/37.02	28,640/37.02	28,640/37.02
Gene classification		Number	
Total genes	133	133	133
Protein-coding genes	88	88	88
tRNA genes	37	37	37
rRNA genes	8	8	8
Genes with introns	16	15	15

Comparison with other *Camellia* species

We compared the lengths of 10 *Camellia* chloroplast genomes, which ranged from 156,585 to 157,121 bp. The GC contents of *C. oleifera* 'Huashuo', 'Huajin' and 'Huaxin' were similar to those of *C. oleifera* samples collected in Hainan. *C. oleifera* from Hainan had the longest chloroplast genome among these four *Camellia* samples, at 156,995 bp, with GC content of 37.31%. The average size of the 10 *Camellia* chloroplast genomes was 156,983 bp. Of the 10 *Camellia* samples, *C. petelotii* had the longest chloroplast genome, and *C. pitardii* the shortest (Table 3). The *C. oleifera* 'Huashuo' chloroplast genome had the smallest IR region (51,906 bp), while *C. sinensis* had the longest (52,180 bp). The *C. oleifera* 'Huashuo' chloroplast genome had the longest SSC region (18,409 bp) and *C. pitardii* had the shortest (18,260 bp) (Table 3). *C. pitardii*

Table 2: Genes identified from the chloroplast genomes of the three *Camellia oleifera* cultivars

Gene categories	Gene groups	Gene names
Genes for photosynthesis	Photosystem I subunits	*psaA, psaB, psaC, psaI, psaJ*
	Photosystem II subunits	*psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbK, psbL, psbM, psbN, psbT, psbZ*
	ATP synthase subunits	*atpA, atpB, atpE, atpF,atpH, atpI*
	Cytochrome b6/f complex subunits	*petA, petB, petD, petG, petL, petN*
	NADH dehydrogenase subunits	*ndhA, ndhB, ndhB-D2, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*
	Large rubisco subunit	*rbcL*
	Small ribosomal subunit proteins	*rps11, rps12, rps12-D2, rps14, rps15, rps16, rps18, rps19, rps2, rps3, rps4, rps7, rps7-D2, rps8*
	Large ribosomal subunit proteins	*rpl14, rpl16, rpl2, rpl2-D2, rpl20, rpl22, rpl23, rpl23-D2, rpl32, rpl33, rpl36*
	RNA polymerase subunits	*rpoA, rpoB, rpoC1, rpoC2*
Other genes	Acetyl-CoA carboxylase	*accD*
	Cytochrome c biogenesis	*ccsA*
	Envelope membrane protein	*cemA*
	Maturase	*matK*
	Protease	*clpP*
	Translation initiation factor	*infA*
Unknown genes	Conserved hypothetical chloroplast reading frame	*orf42, orf42-D2, ycf1, ycf15, ycf15-D2, ycf2, ycf2-D2, ycf3, ycf4*

Table 3: Comparison of the *Camellia* chloroplast genome characteristics

Genome feature	*C. oleifera*'Huashuo'	*C. oleifera*'Huajin'	*C. oleifera*'Huaxin'	*C. sinensis*	*C. petelotii*	*C. azalea*	*C. pitardii*	*C. oleifera*	*C. oleifera* in Hainan	*C. sinensis*cv. Longjing 43
Total length (bp)	156,965	156,975	156,975	157,103	157,121	157,039	156,585	156,971	156,995	157,103
LSC length (bp)	86,650	86,661	86,661	86,646	86,660	86,675	86,213	86,472	86,649	86,646
SSC length (bp)	18,409	18,406	18,406	18,277	18,283	18,282	18,260	18,280	18,298	18,277
IR length (bp)	51,906	51,908	51,908	52,180	52,178	52,082	52,112	52,056	52,048	52,180
GC content (%)	37.29	37.29	37.29	37.31	37.29	37.30	37.34	37.31	37.29	37.31
Total genes	133	133	133	132	132	132	137	132	132	132
Protein genes	88	88	88	87	87	87	89	87	87	87
tRNA genes	37	37	37	37	37	37	40	37	37	37
rRNA genes	8	8	8	8	8	8	8	8	8	8

had the highest GC content (37.34%). The chloroplast genomes of all 10 samples encoded 37 of tRNAs, except for that of *C. pitardii*, which encoded 40 (Table 3).

Repeat sequence analysis

Repeat sequence analysis showed 37 repeats with at least 18 bp per repeat unit in the three *Camellia* chloroplast genomes (Tables S1–S3). These repeats included 19 direct (forward) repeats in *C. oleifera* 'Huashuo' and 'Huaxin' and 18 direct (forward) repeats in *Camellia* 'Huajin'. Fifteen palindrome repeats were detected in *C. oleifera* 'Huashuo' and 'Huajin', and 14 in *C. oleifera* 'Huaxin'. Forward and palindrome repeats were more abundant than reverse and complement repeats in all three *C. oleifera* cultivars. *C. oleifera* 'Huajin' and 'Huaxin' each had two reverse repeats and two complement repeats, whereas *C. oleifera* 'Huashuo' had one reverse repeat (Fig. 4); most were 19–20 bp in length, although *C. oleifera* 'Huashuo' had one 18 bp repeat. In each of the three *Camellia* chloroplast genomes, we also found one repeat each of 23, 24, 26 and 30 bp, four repeats of 38 bp, and two repeats of 42 bp.

SSR analysis

Fifty SSR loci were identified in the chloroplast genomes of *C. oleifera* 'Huashuo' and 'Huaxin', and 51 SSR loci in that of *C. oleifera* 'Huajin' (Tables S4–S6). The maximum length of mononucleotide SSRs among the three *C. oleifera* chloroplast genomes was 17 bp (Fig. 5). These SSR loci were all identified as mononucleotide SSR loci, except for one complicated SSR locus in 'Huashuo'. These mononucleotides repeat units were all type A or T; no G type repeat units were found. These SSR loci contributed to the A/T richness of the three *C. oleifera* chloroplast genomes. These results are similar to those from previous studies of tung tree chloroplast genomes (Li *et al*. 2017). Mononucleotide motif repeat numbers generally range from 10 to 14 bp. In the 'Huashuo', 'Huajin' and 'Huaxin' *C. oleifera* cultivars, the repeat numbers of mononucleotide motifs ranged from 10 to 17, except that 'Huashuo' and 'Huaxin' had no 16-mononucleotide repeats (Fig. 5).

IR expansion/contraction

The IR-SSC and IR-LSC borders of the three *C. oleifera* cultivar chloroplast genomes were compared to those of four other *Camellia* species (*C. petelotii*, *C. sinensis*, *C. oleifera*, and *C. pitardii*) and *N. tabacum*. The *ycf1* pseudogenes were 962 bp long in the three *C.oleifera* cultivars, 1,068 bp in *C. petelotii*, *C. sinensis*, and *C. oleifera*, 1,042 bp in *C. pitardii*, and 1,027 bp in *N. tabacum*. The IRb/SSC borders of eight *Camellia* chloroplast genomes were nested in the *ycf1* gene (962–1068 bp), extending into the IRb region.

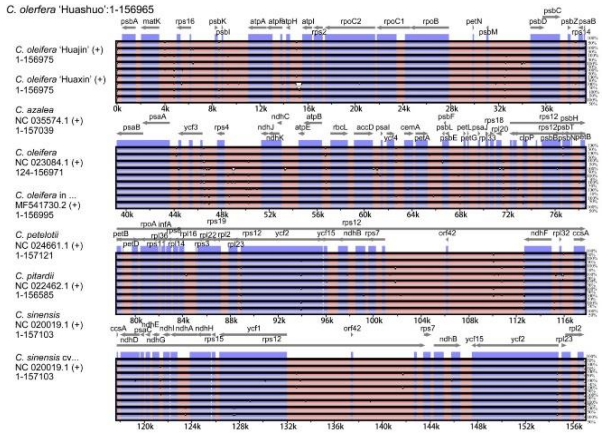

Fig. 3: Comparison of 10 chloroplast genomes using mVISTA. Gray arrows and thick black lines above the alignment indicate gene orientation and inverted repeat (IR) positions, respectively. The vertical scale indicates the percentage identity (50–100%)

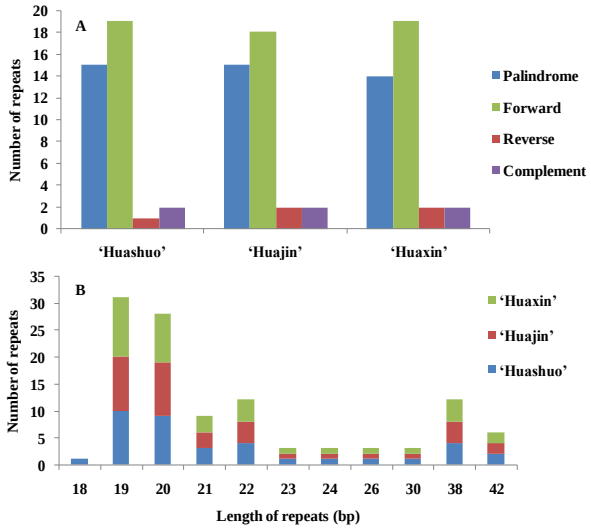

Fig. 4: Repeat sequences in three *C. oleifera* chloroplast genomes. **A**: Repeated sequences in the three *C. oleifera* chloroplast genomes; **B**: Frequencies of four repeat types according to length in the three *C. oleifera* chloroplast genomes

Phylogenetic analysis

The three *C. oleifera* cultivars ('Huashuo', 'Huajin' and 'Huaxin') and *Camellia oleifera* formed a strongly supported monophyly (100%). A sister relationship was revealed among *Camellia oleifera* and the three *C. oleifera* cultivars ('Huashuo', 'Huajin' and 'Huaxin') (100%) (Fig. 7). ML indicated that *C. oleifera* 'Huashuo' was highly supported as a sister to a clade consisting of *C. oleifera* 'Huaxin', *C. oleifera*, and *C. oleifera* 'Huajin'. *C. oleifera* from Hainan was identified as sister to *C. azalea* using bootstrapping (91%) (Fig. S1). *C. oleifera* was suggested to be more closely related to *C. oleifera* 'Huajin' than to *C. oleifera* 'Huaxin' or *C. oleifera* 'Huashuo' (Fig. 7 and S1).

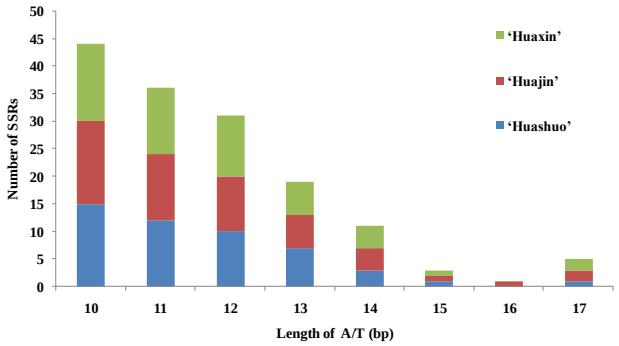

Fig. 5: Distribution of A/T simple-sequence repeats (SSRs) in three *C. oleifera* chloroplast genomes

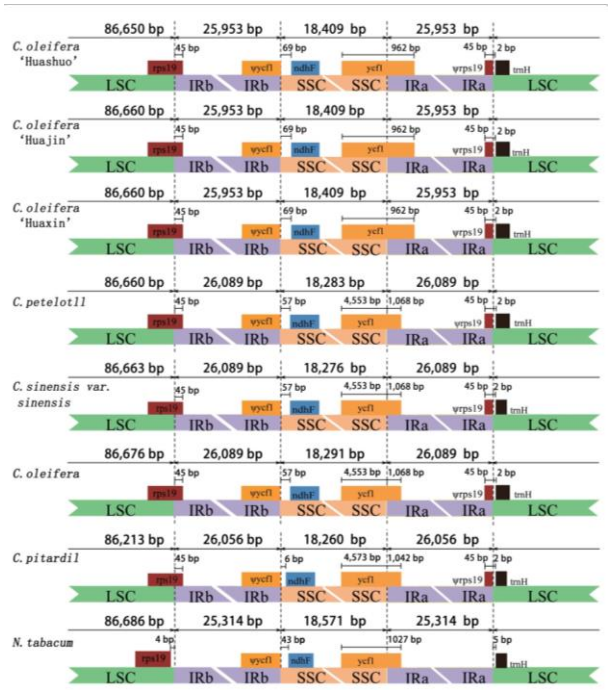

Fig. 6: Comparison of the LSC, IR, and SSC border regions among eight *Camellia* chloroplast genomes

Discussion

The entire chloroplast genomes of three *C. oleifera* cultivars were determined using Illumina HiSeq 4000 Sequencing and third-generation sequencing (PacBio RS II System). Illumina PE (300~500 bp) and PacBio (8~10 kb) libraries were constructed. The obtained sequencing data were mapped using bioinformatics analysis. With Illumina HiSeq sequencing platform to sequence samples results in some low-quality raw data. To ensure that the subsequent analysis was more accurate and reliable, were moved the adapter sequence from the reads, reads with N contents of up to 10%, and those with non-AGCT bases at the 5' end to ensure the accuracy of chloroplast genome assembly. We found that the chloroplast genome sizes were similar among

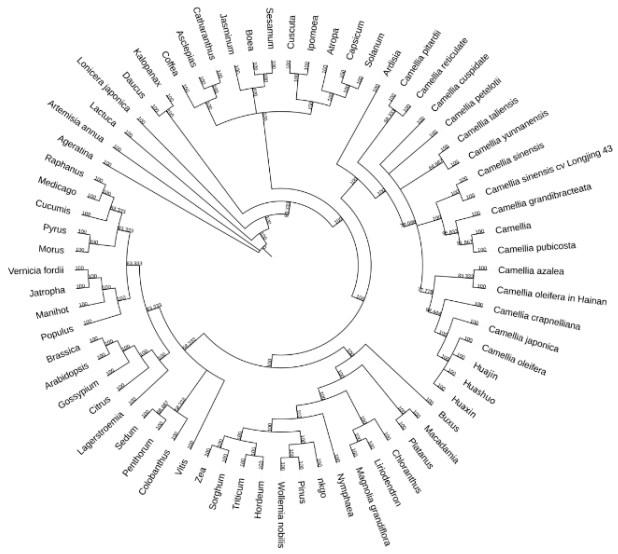

Fig. 7: Phylogenetic tree of 65 taxa based on 50 protein-coding chloroplast genes using the maximum parsimony (MP) default parameters. Bootstrap values (1,000 replications) are shown at the nodes

the three *C. oleifera* cultivars (Table 1), ranging from about 156 to 160 kb, which is typical of *Camellia* species (Wang *et al*. 2017). Whole-chloroplast genome alignment revealed conserved organization and linear gene order among the three *C. oleifera* cultivars and seven other representative *Camellia* chloroplast genomes (Fig. 2 and Table 3). These results are consistent with those reported for herbaceous bamboo (Wang *et al*. 2018) and paleotropical plants (Vieira *et al*. 2015).

The differences in whole-chloroplast genome length are mainly due to differences in the IR region length (Guo *et al*. 2018). The cp genome sequences of *C. oleifera* 'Huashuo', 'Huajin' and 'Huaxin' were then compared with those of seven *Camellia* species using mVISTA. The alignment showed that the 10 cp genomes were conserved, with high gene order. Sequence comparison shows greater divergence is in the LSC and SSC regions than in the IR region, and lower in the coding region than then on-coding region. The 10 *Camellia* chloroplast genomes contained highly differentiated regions in the intergenic spacers. These results are consistent with those for other species (Ni *et al*. 2016; Guo *et al*. 2018; Jian *et al*. 2018). We detected slight variation in the coding regions of some genes including *psbN* and *ycf1* (Fig. 3); variation in the *ycf1* gene has been reported (Jian *et al*. 2018).

Repeat sequences such as SSRs play an important role in the rearrangement and stabilization of cp genome sequences and the copy number variation in different species, even in the same species, characteristics that make them suitable molecular markers for studying genetic diversity (Vieira *et al*. 2014; Su *et al*. 2017). In each of the three *C. oleifera* chloroplast genomes, we found one repeat each of 23, 24, 26, and 30 bp, four repeats of 38 bp, and two repeats of 42 bp. In the three *C. oleifera* chloroplast genomes, most repeats were located in IGS (Tables S1–S3), consistent with results reported by Li *et al*. (2018). In the three *Camellia oleifera* cp genomes, there were 50, 50, and 51 SSR loci at least 10 bp long in *C. oleifera* 'Huashuo', 'Huaxin' and 'Huajin', respectively (Tables S4–S6). Most SSR loci were located in the noncoding regions in the three *C. oleifera* cp genomes. These results are consistent with findings that SSR loci in the cp genome are usually located in IGS regions (Sithichoke *et al*. 2011; Li *et al*. 2017).

Although, the chloroplast genome has a nearly collinear gene order in most land plants, changes in the genome occur in the course of evolution, such as gene loss, sequence inversion, and expansion at the borders of the SSC, LSC, and IR regions (Choi *et al*. 2016; Su *et al*. 2017). The IRs/LSC boundary is a highly informative region for population and phylogenetic studies; for example, the distance between the end edge of *ycf1* and IRb was 257 bp in *Oenothera argillicola* (Gu *et al*. 2018). In all chloroplast genomes examined in this study, the *ndhF* gene was located in the SSC region, 6–69 bp from the IRb/SSC border; it was farthest from the IRb/SSC border in the three *C. oleifera* cultivar chloroplast genomes, and it was nearest to the IRb/SSC border in *C. pitardii*. The *rps19* gene overlapped at IRs in all *Camellia* chloroplast genomes by 45 bp, whereas the *rps19* gene of *N. tabacum* was located in the LSC region, 4 bp from the IRb/LSC border (Fig. 6). Some *rps19* genes are located in the LSC region, some in IR region, especially in monocotyledons, and some at the IRb/LSC border (Wang *et al*. 2016; Li *et al*. 2017). We found that the *rps19* gene positions were exactly the same in seven *Camellia* species, indicating that the *rps19* gene is very stable in *Camellia*. In chloroplast DNA, IRa and IRb are the relatively conserved regions, while expansion and contraction at the borders of IR regions are the main reasons for size variation in chloroplast genomes (Raubeson *et al*. 2007).

Several studies have analyzed phylogenetic relationships within the family Theaceae based on chloroplast coding or non-coding sequences (Yang *et al*. 2013; Huang *et al*. 2014). The chloroplast genome sequence is a useful resource for studying taxonomic status and evolutionary relationships within families (Prince and Parks 2001; Liu *et al*. 2018). The three *C. oleifera* cultivar chloroplast genomes used in this study provide sequence information that can be used in future studies of *Camellia* molecular evolution and phylogeny. To identify the phylogenetic position of *Camellia* within the as terid lineage, we performed multiple sequence alignments using 50 protein-coding genes present in 65 complete chloroplast genome sequences representing 31 orders. Additional chloroplast genomes from *Ginkgo biloba*, *Wollemianobilis* and *Pinus* sp were included as outgroups.

Conclusion

This study analyzed the complete chloroplast genomes of

three *C. oleifera* cultivars ('Huashuo', 'Huajin', and 'Huaxin') that are cultivated widely in China. The genome structure, gene content, and gene number were similar in the three chloroplast genomes and those of other *Camellia* species. Phylogenetic analysis indicated a sister relationship among the three *C. oleifera* cultivars and *C. oleifera*, and *C. oleifera* from Hainan was identified as a sister to *Camellia azalea*. The results provide valuable whole-chloroplast genome information for *Camellia* species that may helpful further phylogenetic analyses of *Camellia* evolutionary relationships and facilitate the genetics and breeding of modern *Camellia*.

Acknowledgments

This work was supported by the Major Projects of Science and Technology Project of Hunan Province (2018NK1030).

Author Contributions

Lingli Wu and Jian'an Li analyzed the results. Ze Li and Xiaofeng Tan prepared plant materials and collected the samples. Fanhang Zhang prepared Fig. 1–4. Yiyang Gu prepared Fig. 5–7. Lingli Wu, Ze Li, and Xiaofeng Tan wrote the main manuscript text. All authors reviewed the manuscript.

References

Antipov D, A korobeynikov, JS McLean, PA Pevzner (2016). HYBRIDSPADES: An algorithm for hybrid assembly of short and long reads. *Bioinformatics* 32:7–13

Ashburner M, CA Ball, JA Blake, D Botstein, JM Cherry (2000). Gene Ontology: Tool for the unification of biology. *Nat Genet* 25:25–29

Beier S, T Thiel, T Münch, U Scholz, M Mascher (2017). MISA-web: A web server for microsatellite prediction. *Bioinformatics* 33:2583–2585

Bobik K, TM Burch-Smith (2015). Chloroplast signaling within, between and beyond cells. *Front Plant Sci* 6; Article 781

Borgstrom E, S Lundin, J Lundeberg (2011). Large scale library generation for high throughput sequencing. *PLoS One* 6; Article e19119

Choi KS, MG Chung, SJ Park (2016). The complete chloroplast genome sequences of three *Veroniceae* species (Plantaginaceae): Comparative analysis and highly divergent regions. *Front Plant Sci* 7; Article 355

Feás X, LM Estevinho, C Salinero, P Vela, MJ Sainz, MP Vázquez-Tato, JA Seijas (2013). Triacylglyceride, antioxidant and antimicrobial features of virgin *Camellia oleifera*, *C. reticulata* and *C. sasanqua* oils. *Molecules* 18:4573–4587

Gao C, DY Yuan, Y Yang, BF Wang, DM Liu, F Zou (2015). Pollen tube growth and double fertilization in *Camellia oleifera*. *J Amer Soc Hortic Sci* 140:12–18

Gu C, B Dong, L Xu, LR Tembrock, S Zheng, Z Wu (2018). The complete chloroplast genome of *Heimia myrtifolia* and comparative analysis within myrtales. *Molecules* 23:846-864

Guo S, L Guo, W Zhao, J Xu, Y Li, X Zhang, X Shen, M Wu, X Hou (2018). Complete chloroplast genome sequence and phylogenetic analysis of *Paeonia ostii*. *Molecules* 23:246-260

Hu JL, SP Nie, DF Huang, L Chang, MY Xie (2012). Extraction of saponin from *Camellia oleifera* cake and evaluation of its antioxidant activity. *Intl J Food Sci Technol* 47:1676–1687

Huang H, C Shi, Y Liu, SY Mao, LZ Gao (2014). Thirteen *Camellia* chloroplast genome sequences determined by high-throughput sequencing: Genome structure and phylogenetic relationships. *BMC Evol Biol* 14; Article 151

Jian HY, YH Zhang, HJ Yan, XQ Qiu, QG Wang, SB Li, SD Zhang (2018). The complete chloroplast genome of a key ancestor of modern roses, *Rosa chinensis* var. *spontanea* and a comparison with congeneric species. *Molecules* 23:389-401

Kang Y, Z Deng, R Zang, W Long (2017). DNA barcoding analysis and phylogenetic relationships of tree species in tropical cloud forests. *Sci Rep* 7; Article 12564

Kazuhiko U, I Hachiro, O Kanji, O Haruo (1984). Nucleotide sequence of *Marchantia polymorpha* chloroplast DNA: A region possibly encoding three tRNAs and three proteins including a homologue of *E. coli* ribosomal protein S14. *Nucl Acids Res* 12:9551–9565

Kurtz S, JV Choudhuri, E Ohlebusch, C Schleiermacher, J Stoye, R Giegerich (2001). Reputer: The manifold applications of repeat analysis on a genomic scale. *Nucl Acids Res* 29:4633–4642

Li X, Y Li, M Zang, M Li, Y Fang (2018). Complete chloroplast genome sequence and phylogenetic analysis of *Quercus acutissima*. *Intl J Mol Sci* 19:2443-2460

Li Z, H Long, L Zhang, Z Liu, H Cao, M Shi, X Tan (2017). The complete chloroplast genome sequence of tung tree (*Vernicia fordii*): Organization and phylogenetic relationships with other angiosperms. *Sci Rep* 7; Article 1869

Li Z, XF Tan, ZM Liu, Q Lin, L Zhang, J Yuan, YL Zeng, LL Wu (2016). In vitro propagation of *Camellia oleifera* abel. using hypocotyl, cotyledonary node, and radicle explants. *HortScience* 51:416–421

Liu X, Y Li, H Yang, B Zhou (2018). Chloroplast genome of the folk medicine and vegetable plant *Talinum paniculatum* (jacq.) gaertn.: Gene organization, comparative and phylogenetic analysis. *Molecules* 23:857-874

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

Magrane M, U Consortium (2011). UniProt Knowledgebase: A hub of integrated protein data. *Database* 2011; Article bar009

Mayor C, M Brudno, JR Schwartz, A Poliakov, I Dubchak (2000). VISTA: Visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* 16:1046–1047

McPherson H, MVD Merwe, SK Delaney, MA Edwards, RJ Henry, E McIntosh, PD Rymer, ML Milner, J Siow, M Rossetto (2013). Capturing chloroplast variation for molecular ecology studies: A simple next generation sequencing approach applied to a rainforest tree. *BMC Ecol* 13; Article 8

Minoru K, G Susumu, K Shuichi, O Yasushi, H Masahiro (2004). The KEGG resource for deciphering the genome. *Nucl Acids Res* 32:277–280

Ni LH, ZL Zhao, HX Xu, SL Chen, G Dorje (2016). The complete chloroplast genome of *Gentiana straminea* (Gentianaceae), an endemic species to the Sino-Himalayan subregion. *Gene* 577:281–288

Palmer JD, RK Jansen, HJ Michaels, CJR Manhart (1988). Chloroplast DNA variation and plant phylogeny. *Ann Missour Bot Gard* 75:1180–1206

Prince LM, CR Parks (2001). Phylogenetic relationships of theaceae inferred from chloroplast DNA sequence data. *Amer J Bot* 88:2309–2320

Qu XJ, H Wang, M Chen, J Liao, J Yuan, GH Niu (2019). Drought stress-induced physiological and metabolic changes in leaves of two oil tea cultivars. *J Amer Soc Hortic Sci* 144:439-447

Raubeson LA, R Peery, TW Chumley, C Dziubek, HM Fourcade, JL Boorem, RK Jansen (2007). Comparative chloroplast genomics: Analyses including new sequences from the angiosperms *Nupharadvena* and *Ranunculus macranthus*. *BMC Genomics* 8; Article 174

Ravi V, JPK hurana, AK Tyagi, P Khurana (2006). The chloroplast genome of mulberry: Complete nucleotide sequence, gene organization and comparative analysis. *Tree Genet Genomics* 3:49–59

Sithichoke T, U Pichahpuk, S Duangjai (2011). Characterization of the complete chloroplast genome of *Hevea brasiliensis* reveals genome rearrangement, RNA editing sites and phylogenetic relationships. *Gene* 475:104–112

Song Y, Y Chen, J Lv, J Xu, S Zhu, M Li, N Chen (2017). Development of Chloroplast Genomic Resources for *Oryza* Species Discrimination. *Front Plant Sci* 8; Article 1854

Stamatakis A (2006). Raxml-vi-hpc: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690

Su YH, SC Kyeong, OY Ki, OL Hyun, SC Kwang, TS Jong (2017). Complete chloroplast genome sequences and comparative analysis of *Chenopodium quinoa* and *C. album*. *Front Plant Sci* 8; Article 1696

Tan XF, TQ Guan, J Yuan (2018). Investigation and research of upgrading and building a hundred billion yuan oil tea industry in hunan province. *Nonwood For Res* 36:1–4

Tan XF, DY Yuan, F Zou, J Yuan, P Xie, Y Su, Y Wang, DT Yang, JT Peng (2012). An elite variety of oil tea: *Camellia oleifera* 'Huaxin'. *Sci Silv Sin* 48:170–171

Tan XF, DY Yuan, J Yuan, F Zou, P Xie, Y Su, DT Yang, JT Peng (2011). An elite variety of oil tea: *Camellia oleifera* 'Huashuo'. *Sci Silv Sin* 47:184–209

Tatusov RL, ND Fedorova, JD Jackson, AR Jacobs, B Kiryutin, EV Koonin, DM Krylov, R Mazumder, SL Mekhedov, AN Nikolskaya, SB Rao, S Smirnov, AV Sverdlov, S Vasudevan, YI Wolf, JJ Yin, DA Natale (2003). The COG database: An updated version includes eukaryotes. *BMC Bioinform* 4; Article 41

Vieira LDN, KGD Anjos, H Faoro, HPDF Fraga, TM Greco, FDO Pedrosa, EMD Souza, M Rogalski, RFD Souza, MP Guerra (2015). Phylogenetic inference and SSR characterization of tropical woody bamboos tribe Bambuseae (Poaceae: Bambusoideae) based on complete plastid genome sequences. *Curr Genet* 62:1–11

Vieira LDN, H Faoro, M Rogalski, HPDF Fraga, RLA Cardoso, EMD Souza, FBDO Pedrosa, RO Nodari, MP Guerra (2014). The complete chloroplast genome sequence of *Podocarpus lambertii*: Genome structure, evolutionary aspects, gene content and SSR detection. *PLoS One* 9; Article e90618

Wang G, Y Luo, N Hou, LX Deng (2017). The complete chloroplast genomes of three rare and endangered camellias (*Camellia huana*, *C. liberofilamenta* and *C. luteoflora*) endemic to southwest china. *Conserv Genet Resour* 9:583–589

Wang L, TN Wuyun, HY Du, DP Wang, DM Cao (2016). Complete chloroplast genome sequences of *Eucommia ulmoides*: Genome structure and evolution. *Tree Genet Genomics* 12; Article 12

Wang W, S Chen, X Zhang (2018). Whole-genome comparison reveals divergent ir borders and mutation hotspots in chloroplast genomes of herbaceous bamboos (Bambusoideae: Olyreae). *Molecules* 23:1537-1556

Wang YH, Y Zhang, R Wang, P Liang, F Liu, LC Wu (2019). Research on comprehensive evaluation of Camellia oil quality based on principal component analysis. *J Centr South Univ For Technol* 39:45–51

Wu FH, MT Chan, DC Liao, CT Hsu, YW Lee, H Daniell, MR Duvall, CS Lin (2010). Complete chloroplast genome of *Oncidium* Gower and evaluation of molecular markers for identification and breeding in Oncidiinae. *BMC Plant Biol* 10; Article 68

Wu LL, JA Li, YY Gu, FH Zhang, L Gu, XF Tan, MW Shi, (2020). Effect of chilling temperature on chlorophyll florescence, leaf anatomical structure, and physiological and biochemical characteristics of two *Camellia oleifera* cultivars. *Intl J Agric Biol* 23:777–785

Wyman SK, RK Jansen, JL Boore (2004). Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20:3252–3255

Xu C, WP Dong, WQ Li, YZ Lu, XM Xie, XB Jin, JP Shi, KH He, ZL Suo (2017). Comparative analysis of six lagerstroemia complete chloroplast genomes. *Front Plant Sci* 8; Article 15

Yang JB, SX Yang, HT Li, J Yang, DZ Li (2013). Comparative chloroplast genomes of *Camellia* species. *PLoS One* 8; Article e73053

Yuan DY (2012). An elite variety: *Camellia oleifera* 'Huajin'. *Sci Silv Sin* 48:170

Zeng YL, XF Tan, L Zhang, HX Long, BM Wang, Z Li, Z Yuan (2015). A fructose-1,6-biphosphate aldolase gene from *Camellia oleifera*: Molecular characterization and impact on salt stress tolerance. *Mol Breed* 35:1–17

Zhang Y, GD Huang, RW Li, YL Mo, Q Huang (2017). Research on extraction methods of chloroplast DNA in *Mangifera* L. *Nonwood For Res* 35:50–54

Zhu WF, CL Wang, F Ye, HP Sun, CY Ma, WY Liu, F Feng, M Abe, T Akihisa, J Zhang (2018). Chemical constituents of the seed cake of *Camellia oleifera* and their antioxidant and antimelanogenic activities. *Chem Biodivers* 15; Article e1800137