Characterization of the Complete Chloroplast Genome Sequences and Phylogenetic Relationships of Four Oil-Seed Camellia spp. and related taxa

Huihua Luo† 1, Boyong Liao† 1, Yongjuan Li 1, Runsheng Huang 1, Kunchang Zhang 1, Longyuan Wang 1, Jing Tan 1,Yuzhou Lv 2, Can Lai 1*, Yongquan Li 1*

[1] College of Horticulture and Landscape Architecture, Zhongkai University of Agriculture and Engineering, Guangzhou, Guangdong 510220, China; zk_lhh2022@126.com (H.L.); liaoby05@126.com (B.L.); liyongjuan@zhku.edu.cn (Y.J.L.); runsheng_huang@foxmail.com (R.H.); 1007849940@qq.com (K.Z.); wanglongyuan@zhku.edu.cn (L.W.); 670108097@qq.com (J.T.); laican@zhku.edu.cn (C.L.); yongquanli@zhku.edu.cn (Y.Q.L.)

[2] State-owned Xiaokeng Forest Farm in Qujiang District of Shaoguan City, Shaoguan, Guangdong 512100, China; 2972913466@qq.com (Y.Z.L.)

[*] Correspondence: laican@zhku.edu.cn (C.L.); yongquanli@zhku.edu.cn (Y.Q.L.)

[†] These authors contributed equally to this work.

Keywords: *Camellia*; Oil-seed; Chloroplast genome structure; Phylogenetic relationship; hybridization and chromosome polyploid.

Abstract

Some species in the Sect. *Oleifera* of the genus *Camellia* L. known as oil-seed camellia because of their high oil content and economic value. Additional studies aimed at clarifying the phylogenetic relationships and chloroplast genomes of *Camellia* species are needed to hybridization, as well as improve the breeding, selection and interspecific hybridization of *Camellia* species. The complete chloroplast genomes (cpDNA) of the four oil-seed camellia species *C. semiserrata*, *C. meiocarpa*, *C. suaveolens*, and *C. osmantha* were resequenced to clarify their interspecific relationships. These cpDNA had typical tetrad structures, and they were highly conserved in various structural features. The total lengths of the cpDNA ranged from 156,965 to 157,018 bp, and 134 genes were annotated, including 88 protein-coding genes, 37 transfer RNA genes, and 8 messenger RNA genes. The average GC content of these genomes was 37.3%. The codons with the highest and lowest codon usage bias were UUA (which codes for leucine) and AGC (which codes for serine), respectively. The number of simple sequences repeats of the four *Camelia* species ranged from 38 to 40. Mononucleotide repeats were the most common repeat type, followed by tetranucleotide, trinucleotide, and hexanucleotide repeats. Our phylogenetic analysis of cpDNA, coupled with the results of previous ploidy analyses and artificial interspecific hybridization, revealed that *C. semiserrata* was most closely related to *C. azalea*, *C. suaveolens* was most closely related to *C. gauchowensis*, *C. osmantha* was most closely related to *C. vietnamensis*, and *C. meiocarpa* was most closely related to *C. oleifera*. The phylogenetic relationships between oil-seed camellia species with high oil content and economic value were characterized. Our analysis of the cpDNA provided new insights that will aid the use of artificial distant hybridization in camellia breeding programs.

Introduction

Oil-seed camellia plants have become economically important woody oil plants in China in recent years. Some plants in the genus *Camellia* L. (family Theaceae) are generally referred to as oil-seed camellia because of their high oil content and economic value (FRPS,1998; State owned forest farm and forest seedling work station of the State Forestry Administration, 2016). Oil-seed camellia plants are unique woody oil plants and the fourth largest in the world in terms of production after *Olea europaea* L., *Elaeis guineensis* Jacq., and *Cocos nucifera* L.. Camellia oil is highly nutritious, and its unsaturated fatty acid content is up to 90% higher than that of other common edible oils on the market. Oil from camellia seeds has been shown to be effective in preventing cardiovascular and cerebrovascular diseases. Camellia oil is also considered a healthy edible vegetable oil by the Food and Agriculture Organization. It has various uses in industry, such as the production of cosmetic products and medicine. In addition, the potential applications of these residues from the oil pressing process are manifold, and additional studies will likely broaden the uses of these residues (Li et al., 2011).

Oil-seed camellia plants have been cultivated for approximately 2, 300 years (Wang et al., 2020). China contributes approximately 95% of the world's production of *Camellia* plants and 90% of the world's production of *Camellia* seeds. The largest oil-tea cultivating areas in China are in the following provinces of Hunan, Jiangxi, Guangxi, Zhejiang, Fujian, and Guangdong. Oil-seed camellia is also grown in 1, 537 counties (cities) in 14 other provinces in China. The main oil-seed camellia tree species cultivated in China are *C. oleifera*, *C. meiocarpa*, *C. gauchowensis*, and *C. semiserrata*. Edible oil is the main product of *Camellia* cultivation.

Although various classifications of the genus *Camellia* have been proposed, these classifications have mostly been based on morphological characters and molecular information from DNA sequences and chloroplast sequences (Chang, 1981). Given that hybridization and polyploidization are frequent in this group, traditionally used morphological indexes are likely affected by various environmental factors such as topography, soil, and climate. Trees reconstructed based on a few genes and chloroplasts often show inconsistent or reticulate evolution. This poses a major challenge to taxonomic and phylogenetic analyses of the genus *Camellia* L. and has greatly limited progress in our understanding of the classification of *Camellia* L. (Min et al., 1996; Hong, 1981). The phylogenetic relationships among *Camellia* members remain unclear. Other non-morphological sources of data are needed to clarify their evolutionary relationships, such as, whole genome information, chromosome variation and the utility of artificial interspecific hybridization (Zhang et al., 2022; Yu et al., 2022; Ye et al., 2021; Zhong et al., 2020; Chang et al., 2016; Zhou et al., 2001).

The chloroplast is an important site for energy conversion and photosynthesis in green plants. Chloroplast genomes (cpDNA) are one of the three major genetic systems in plants. They have been widely used in evolutionary studies because of their low nucleotide substitution rates, uniparental inheritance, conserved structure, and low molecular weight (Tang et al., 2022; Tian et al., 2021; Zhu et al., 2022). The cpDNA of these species have been resequenced several times to clarify their relationships, such as the newly described oil-seed species of *C. osmantha* (Ma et al., 2012; Liu et al., 2021). Here, we used next-generation high-throughput sequencing to assemble, annotate, and characterize the cpDNA of four *Camellia* species. We generated more chloroplast genomic resources for oil-seed *Camellia* to characterize the structure of cpDNA and clarify the relationships among these four oil-seed camellia species within the genus *Camellia* L. Our findings provide new insights into chromosome variation and the utility of artificial interspecific hybridization for camellia breeding. The results of our study will also aid future studies aimed at genetically improving the oil production of several oil-seed camellia species.

Materials And Methods

Experimental Materials and Sequencing

Seeds of *C. semiserrata*, *C. meiocarpa*, *C. suaveolens*, and *C. osmantha* were collected from Zhaoqin, Meizhou, Lechang, and Nanning in southern China in 2013 (**Table 1**). The seedlings of four camellia species were planted in Xiaokeng state-owned forest farm (24°15' N, 113°35' E) in Qujiang District, Shaoguan City, Guangdong Province, China in 2015. Young leaves of four species without signs of pests and disease were collected in bags filled with dried silica gel prior to transport to the laboratory in 2020. All four species were identified by the Sun Yat-sen University Herbarium. The remaining silica gel-dried young leaves and vouchers were deposited in the Zhongkai University of Agriculture and Engineering Herbarium for subsequent studies.

A modified CTAB method was used to extract total DNA (Rogers & Bendich, 1989). The Illumina NovaSeq platform was used to construct, quality, and paired-end sequence (2×150bp) the DNA libraries using the Illumina NovaSeq platform. Information on the sequencing data generated by Guangzhou Nuosai Biotechnology Co., Ltd. and used for further assembly and annotation is shown in Table 1. The paired-end read sequences of the four species were submitted to the National Center for Biotechnology Information (NCBI) database (BioProject ID: PRJNA931566).

Table 1. Sample sequencing data information.

Species	Voucher NO.	Seed sample collection sites	Seed collection year	Raw base (BP)	Raw Q30(%)
C. semiserrata Chi.	ZKU2020032	Guangning District, Zhaoqin City, Guangdong Province, China	2013	11, 788, 927, 200	90.56; 87.11
C.meiocarpa Hu.	ZKU2020035	Xingning District, Meizhou City,Guangdong Province, China	2013	9, 157, 311, 900	91.22; 86.76
C. suaveolens Ye.	ZKU2020051	Lechang City, Guangdong Province, China	2013	8, 529, 454, 200	91.67; 88.99
C. osmantha Ye.	ZKU2020159	Guangxi Forestry Research Institute, Nanning, China	2013	10, 930, 712, 700	91.53; 87.79

Assembly and Annotation of cpDNA

The cpDNA were extracted and assembled from the original data using GetOrganelle V1.7.5.3 software (Jin et al., 2020). After the cpDNA were assembled, the chloroplast genome loop with a typical tetrad structure was concatenated and visualized using Bandage software (Wick et al., 2015). The chloroplast genome of *C. vietnamennsis* (NC_060778.1) was used as the reference genome, and the sequenced genomes were annotated using PGA software (Qu et al., 2019). Geneious 9.0.2 software was used to visualize the annotated sequences (Kearse et al., 2012), and manual adjustments were made to improve the sequences. The annotation results were submitted to the National Center for Biotechnology Information (NCBI) under the accession numbers ON367462, ON418964, ON418963, and ON418965. The online software OGDRAW (Greiner et al., 2019) was used to map the cpDNA.

Sequence Alignment Analysis of cpDNA

The four cpDNA were assembled and uploaded to NCBI. The cpDNA of *C. vietnamensis*, *C. crapnelliana*, *C. oleifera*, *C. chekiangoleosa*, *C. gauchowensis*, and *C. kissii* were downloaded to conduct analyses. The GenBank accession numbers were NC060778, KF753632, KY406750, NC037472, NC053541, and NC053915, respectively (**Supplementary table S1**). ShufflemVISTA software (https://genome.lbl.gov/vista/mvista/submit.shtml) was used to align the cpDNA of the different species using the alignment subprogram Shuffle-LAGAN with global pair-wise alignment of the finished sequences (Chris et al., 2000). The online software IRscope (https://irscope.shinyapps.io/irapp/) (Amiryousefi et al., 2018) was used to make comparisons of the annotated cpDNA of *C. semiserrata*, *C. meiocarpa*, *C. suaveolens*, and *C. osmantha* with the downloaded sequences in the inverted repeat (IR) regions of the cpDNA, including regions of contractions and expansions. The molecular markers of the cpDNA for the above 10 *Camellia* species were developed using DnaSP6 software (Rozas et al., 2017).

Repeat Sequence Analysis and Codon Bias Analysis

The online software MISA (https://webblast.ipk-gatersleben.de/misa/) (Beier et al., 2017) was used to localize repetitive elements in the cpDNA of *C. semiserrata*, *C. meiocarpa*, *C. suaveolens*, *C. osmantha*, and *C. vietnamensis*. To detect simple sequence repeats (SSRs), the minimum number of repetitions for mononucleotide, dinucleotide, trinucleotide, tetranucleotide, and hexanucleotide repeats was set to 12, 6, 4, 3, 3, and 3, respectively. The online software REPuter (Kurtz et al., 2001) was used to identify interspersed nuclear elements. Four types of repeat sequences were detected, including forward repeat (F), reverse repeat, palindromic repeat (P), and complementary repeat sequences. The minimum repeat size was set at 30 bp, and the minimum repeat distance was set at 3, respectively. Statistical analyses were conducted using Microsoft Excel 2019, and figures were made using Origin 2022.

The protein-coding genes (CDS) of *C. semiserrata*, *C. meiocarpa*, *C. suaveolens*, and *C. osmantha* were extracted using Geneious 9.0.2. To reduce the errors caused by short sequences, duplicate sequences and sequences less than 300 bp in length were removed from these coding sequences. Next, coding sequences that start with an ATG and end with a TAA, TGA, and TAG were selected. Finally, CodonW 1.4.2 software was used to conduct relative synonymous codon usage (RSCU) analysis on each of the CDS of each species to characterize patterns of codon bias (Sharp and Li, 1987).

Phylogenetic Analysis

The cpDNA of 29 *Camellia* species and two outgroup species (*Schima superba* and *Tutcheria pingpiensis*) were downloaded from the NCBI database to construct phylogenetic trees with the resequenced cpDNA of *C. semiserrata*, *C. meiocarpa*, *C. suaveolens*, and *C. osmantha* (**Supplementary Table S1**). Geneious 9.0.2 software was used to select sequences of regions from the large single-copy (LSC), IR, small single-copy (SSC), and 68 CDS. A multiple sequence alignment of the whole genome and the selected sequences was conducted using MAFFT v7.308 software (Kazutaka et al., 2017). A phylogenetic tree was constructed for each sequence. Phylogenetic trees were constructed using MrBayes ver. 3.2.7a(http://nbisweden.github.io /MrBayes/index.html) and RAxML software with 1, 000 bootstrap replicates (Ronquist et al., 2012; Stamatakis, 2014). FigTree 1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/) was used to build the Bayesian and maximum likelihood (ML) tree, and the iTOL online tool (https://itol.embl.de/) was used to visualize the phylogenetic relationships (Letunic and Bork, 2021).

Results

Basic Characteristics of the cpDNA of Four Species

The cpDNA of the four *Camellia* species exhibited a typical tetrad structure (**Figure 1 and Table 2**), with a total length ranging from 156, 965 to 157, 018 bp, consisting of an LSC region (ranging from 866, 647 to 86, 656 bp), an SSC (ranging from 18, 282 to 18, 408 bp), and a pair of IR regions (ranging from 25, 954 to 26, 042 bp). The lengths of the four cpDNA only varied by 53 bp. The cpDNA of *C. semiserrata* was the longest, and that of *C. meiocarpa* was the shortest. The LSC region was the least variable region in the cpDNA of the four *Camellia* species (only varying by 9 bp), whereas the SSC region was the most variable region among the four *Camellia* species (varying by 126 bp). The average GC content of the four chloroplast genomes was 37.3%. The GC content of the LSC region was 35.3%, and that of the SSC region ranged from 30.5 to 30.6%. The GC content of the IR region (43.0%) was higher than that of the LSC and SSC regions. A total of 134 genes were identified in the four cpDNA, including 88 CDS, 37 transfer RNA (tRNA) genes, and 8 messenger RNA (mRNA) genes. The pseudogene ycf1 was located in the boundary region between the IRB and SSC regions, and other incomplete copy of *ycf1* was also located at the boundary between the SSC and IRA regions. The cpDNA of the four *Camellia* species are highly conserved given that they all possess the basic structural features of cpDNA.

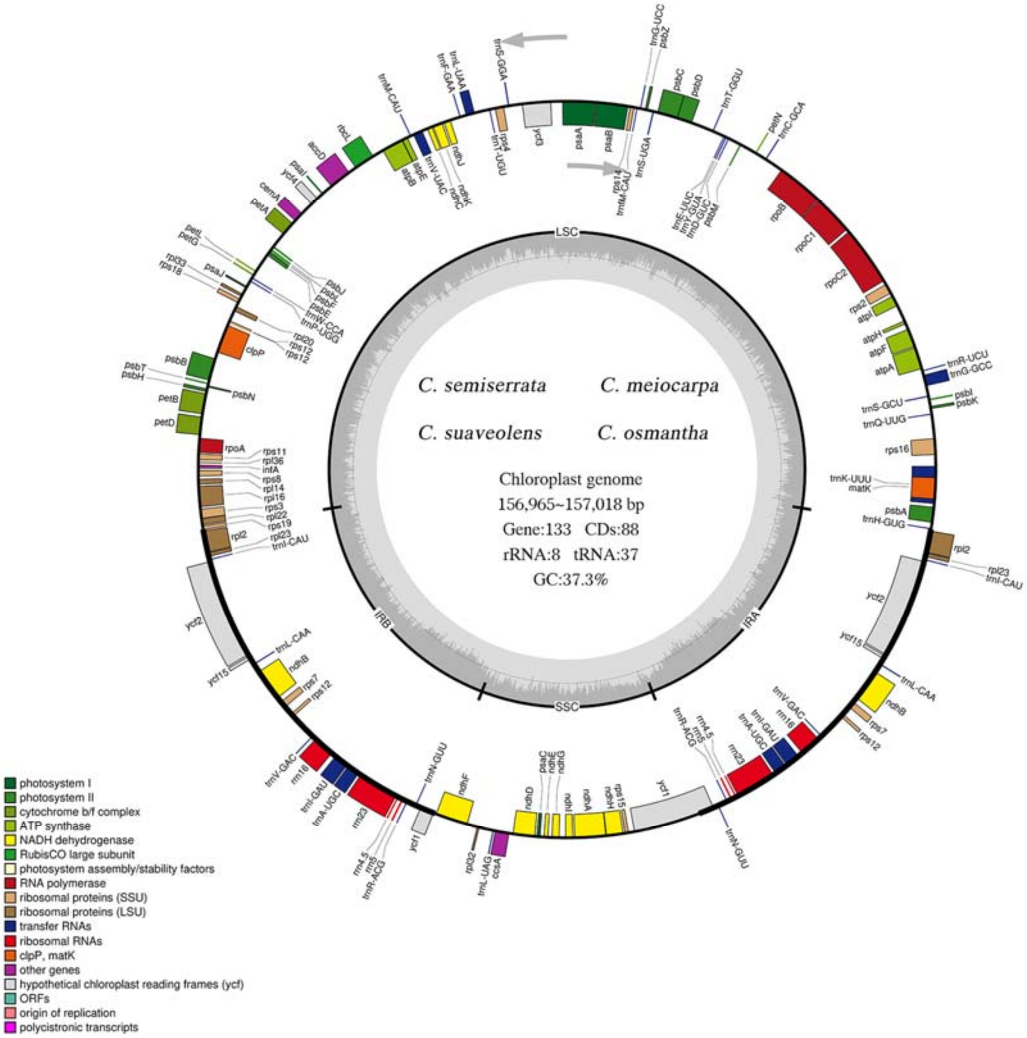


Figure 1. Gene map of the cpDNA of four *Camellia* species. The genes inside the circle are transcribed clockwise, and the genes outside is transcribed counterclockwise. Genes with different functions are color-coded. The darker gray in the center circle indicates the GC content, and the lighter gray indicates the AT content.

Table 2. Analysis of the cpDNA characteristics of four *Camellia* species.

Species	*C. semiserrata**	*C. meiocarpa**	*C. suaveolens**	*C. osmantha**	*C. vietnamensis*
Total cp genome size (bp)	157, 018	156, 965	157, 003	156, 981	156, 999
LSC region (bp)	86, 652	86, 649	86, 656	86, 647	86, 652
IR region (bp)	26, 042	25, 954	26, 025	26, 025	26, 025
SSC region (bp)	18, 282	18, 408	18, 297	18, 284	18, 297
GC content /%	37.3	37.3	37.3	37.3	37.3
GC content in LSC region (%)	35.3	35.3	35.3	35.3	35.3
GC content in IR region (%)	43.0	43.0	43.0	43.0	43.0
GC content in SSC region (%)	30.6	30.5	30.6	30.5	30.5

gene	134	134	134	134	134
CDS	88	88	88	88	87
rRNA	8	8	8	8	8
tRNA	37	37	37	37	37

Note："*" indicates the species resequenced in this study.

Excluding the one pseudogene, 98 out of 133 genes were unique, including 75 CDS and 23 tRNA genes. The other 35 genes were located in the IR regions, including 13 CDS (*ycf1*, *rps7* *2, *ndhB**2, *ycf15* *2, *ycf2* *2, *rpl23* *2, and *rpl2* *2), 8 ribosomal RNA (rRNA) genes, and 14 tRNA genes. A total of 45 genes were involved in photosynthesis, 75 genes were involved in self-replication, 6 genes encoded other proteins, and 7 genes had unknown functions. Sixteen genes had one intron, and two genes had two introns (**Table 3**).

Table 3. List of genes found in the cpDNA of four *Camellia* species.

gene category	gene group	gene name
Photosynthesis	photosystem I	*psaA, epsaB, psaC, psaI, psaJ*
	photosystem II	*psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*,*psbN* , *psbT*, *psbZ*
	NADH dehydrogenase	*ndhA**, *ndhB**(2), *ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*
	cytochrome b/f complex	*petA*, *petB**, *petD**, *petG*, *petL*, *petN*
	ATP synthase	*atpA*, *atpB*, *atpE*, *atpF**, *atpH*, *atpI*
	rubisco	*rbcL*
Self-replication	large ribosomal subunit	*rpl14*, *rpl16**, *rpl2**(2), *rpl20*, *rpl22*, *rpl23*(2), *rpl32*, *rpl33*, *rpl36*
	small ribosomal subunit	*rps11*, *rps12**(2), *rps14*, *rps15*, *rps16**, *rps18*, *rps19*, *rps2*, *rps3*, *rps4*, *rps7*(2), *rps8*
	RNA polymerase	*poA*, *rpoB*, *rpoC1**, *rpoC2*
	ribosomal RNA	*rrn16*(2), *rrn23*(2), *rrn4.5*(2), *rrn5*(2)
	transfer RNA	r*nA-UGC**(2), *trnC-GCA*, *trnD-GUC*, *trnE-UUC*, *trnF-GAA*, t*rnfM-CAU*, *trnG-GCC**, *trnG-UCC*, *trnH-GUG* , *trnI-CAU*(2), *trnI-GAU**(2), *trnK-UUU**, *trnL-CAA*(2), *trnL-UAA**, *trnLUAG*, *trnM-CAU*(2), *trnN-GUU* (2), *trnP-UGG*, *trnQ-UUG*, *trnR-ACG*(2), *trnR-UCU*, *trnS-GCU*, *trnS-GGA*, *trnS-UGA*, *trnT-GGU*, *trnT-UGU*, *trnV-GAC* (2), *trnV-UAC* *, *trnW-CCA*, *trnY-GUA*
Other genes	maturase	*matK*
	protease	*clpP***
	carbon metabolism	*cemA*
	fatty acid synthesis	*accD*
	cytochrome C synthesis	*ccsA*
	Translation initiation factor	*infA*
Unkown functional genes	hypothetical chloroplast open reading frames	*ycf1*(2), *ycf2* (2), *ycf3***, *ycf4*, *ycf15* (2)

Note: “*” and “**” indicate that the gene contains one or two introns, respectively; data in parentheses indicate copy number.

Contractions and Expansions of the IR Boundary

IR boundary analysis of 10 *Camellia* species, including the four *Camellia* species in this study (**Figure 2**), revealed that the structure and sequences of the IR boundary regions of *C. semiserrata*, *C. meiocarpa*, *C. suaveolens*, *C. osmantha*, and *C. vietnamensis* were similar. The genes of *rpl22*, *rps19*, *rpl2*, *ycf1*, *ndhF*, and *trnH* were mainly located near the IR/LSC and IR/SSC boundaries of the cpDNA for these 10 *Camellia* species. The *rps19* gene crossed the LSC/IRB boundary in nine of these species. The IR region of *C. chekiangoleosa* has undergone a contraction. Consequently, the *rpl2* gene has expanded to the LSC region and crosses the LSC/IRB boundary. The *rpl2* gene in *C. kissii* and *C. gauchowensis* was not present in the LSC/IRB boundary. In *C. semiserrata*, *C. meiocarpa*, *C. suaveolens*, *C.osmantha*, and *C. vietnamensis*, *ycf1* is a pseudogene in the SSC/IRA boundary, and there is also an incomplete copy of *ycf1* in the SSC/IRB boundary region. The *ycf1* gene in *C. meiocarpa* is located 1 bp away from the SSC/IRA boundary. And the *ycf1* gene in the other four species is located 106 bp away from the SSC/IRA boundary. The locations of *ndhF* and *ycf1* in the SSC/IRB and SSC/IRA boundary regions vary in *C. kissii*, *C. oleifera*, *C. crapnelliana*, and *C. chekiangoleosa*. This is mainly associated with the contraction and expansion of the IR and SSC regions. The locations of *rpl2* and *trnH* in the LSC/IRA boundary region are consistent in other plants, with the exception of *C. gauchowensis* and *C. crapnelliana*, which did not possess the *rpl2* gene.

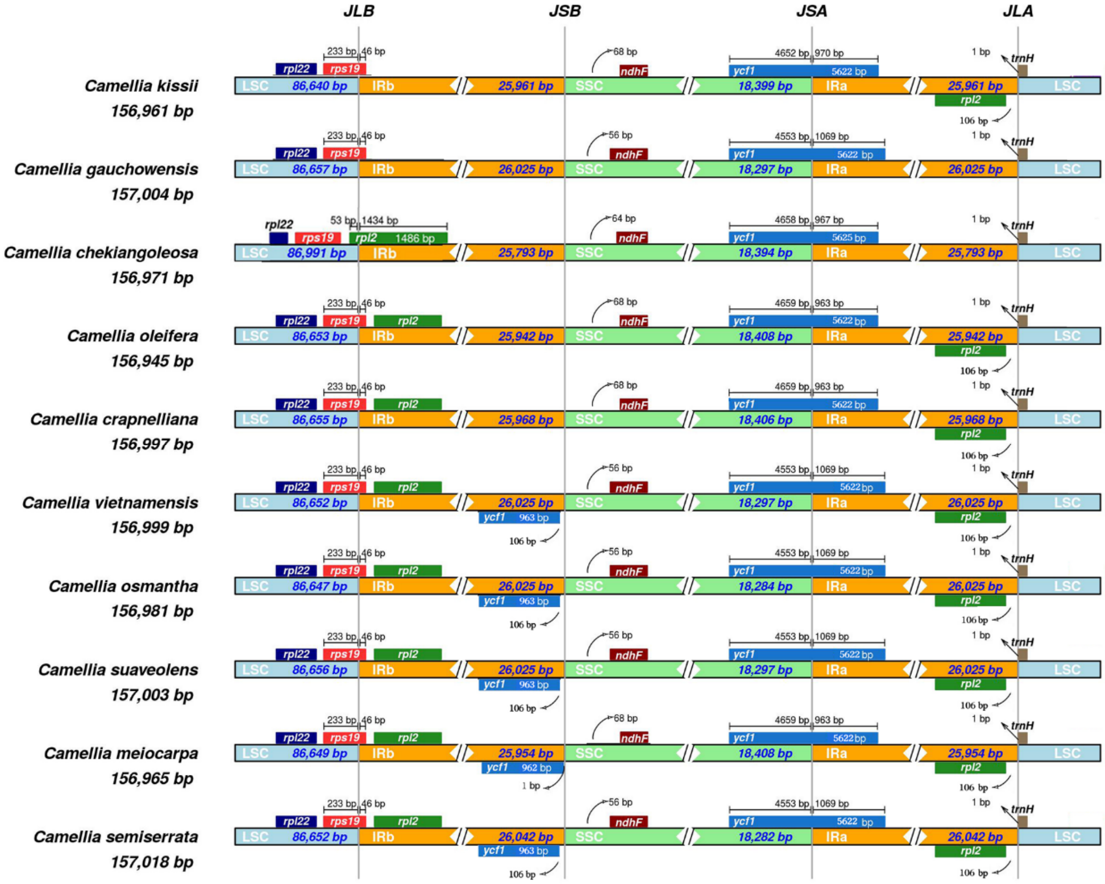


Figure 2. Comparison of IR/SC boundary regions of 10 *Camellia* species. LSC, Large single-copy; SSC, Small single-copy; IRA and IRB, inverted repeats. JLB, junction between LSC and IRB; JSB, junction between SSC and IRB; JSA, junction between SSC and IRA; JLA junction between LSC and IRA.

Molecular Marker Detection

The cpDNA of 10 *Camellia* species, including the four focal *Camellia* species in this study, were compared with mVISTA software using *C. vietnamensis* as a reference (**Figure 3**). The results showed that sequence similarity in the coding region was high. However, sequence similarity of the non-coding region was low. The LSC region was the most variable region across the cpDNA, followed by the SSC, IRB, and IRA regions. Thus, variation in the IR region is low, which suggests that it is evolutionarily conserved. The most conserved genes were rRNA genes, as no significant variation was observed in rRNA genes among cpDNA.

Molecular markers were developed for these 10 *Camellia* species (**Figure 4**). Five highly variable gene spacers or genes were identified, including *trnQ-UUG—trnG-GCC*, *petN-pstM*, and *trnL-UAA—ndhJ* in the LSC region, *ycf15—ndhB* in the IR region, and *ycf1* in the SSC region. In the *trnQ-UUG—trnG-GCC* region from the LSC region, the pi value was highest from 9, 296 to 9, 905 bp (0.01737). This region has a total of 21 mutated sites and, GC content of 30.7%. It was thus the most highly mutated region in the entire cpDNA. In the *ycf15—ndhB* region of the IR region, pi was highest in the 96,233–96,837-bp, 96,438–97,037-bp, and 147,251–147,850-bp regions, all of which were approximately 0.00437. There were total of six mutated sites, and the GC content was 42.8% in this region. In the ycf1 gene in the SSC region, pi was highest in the 128,406–129,005-bp region (0.00463) with nine mutated sites, and a GC content of 23.7%. These regions with high variability can be used as DNA barcodes for species identification.

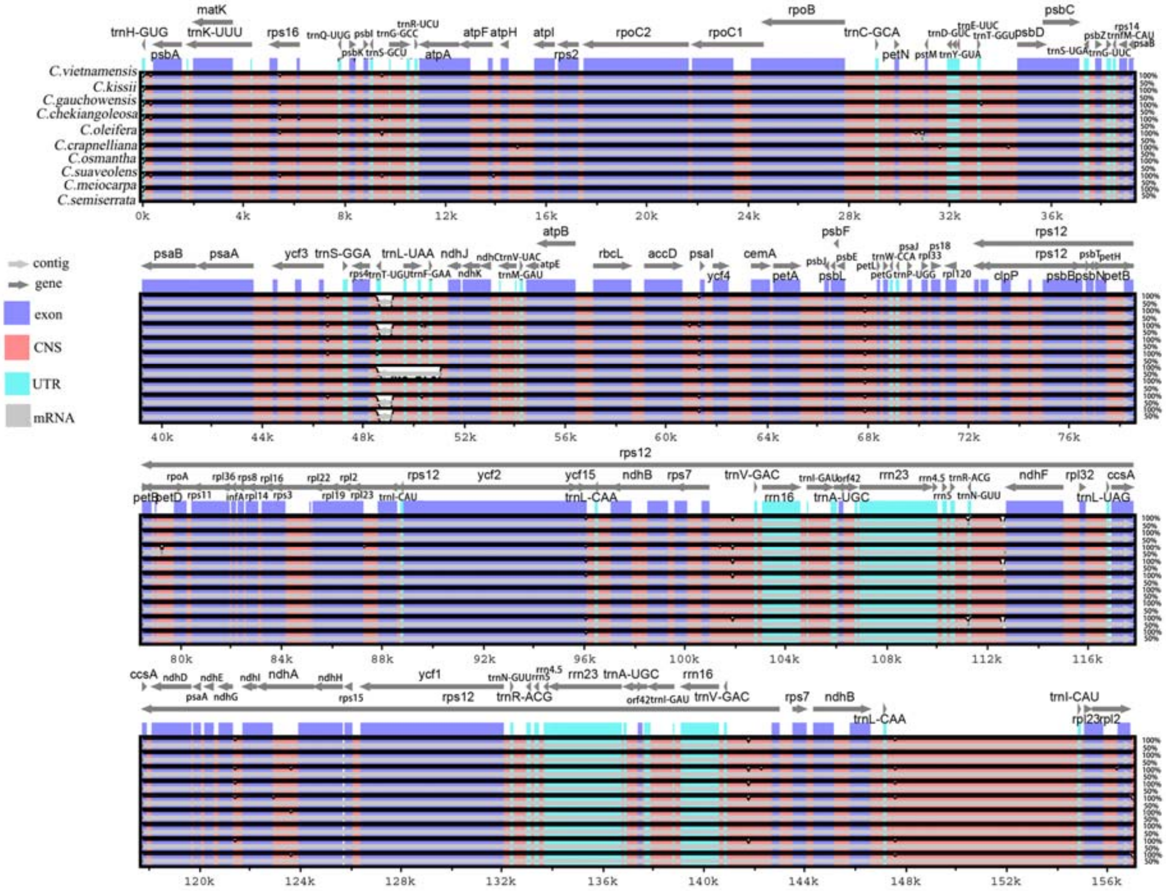


Figure 3. Comparison of the cpDNA of 10 *Camellia* species. The X-axis indicates aligned base sequences, and the Y-axis indicates percent pairwise identity within 50-100%. Grey arrows represent genes and their directions. Blue boxes indicate exon regions, light blue boxes indicate regions encoding RNA genes, and red boxes indicate non-coding sequences.

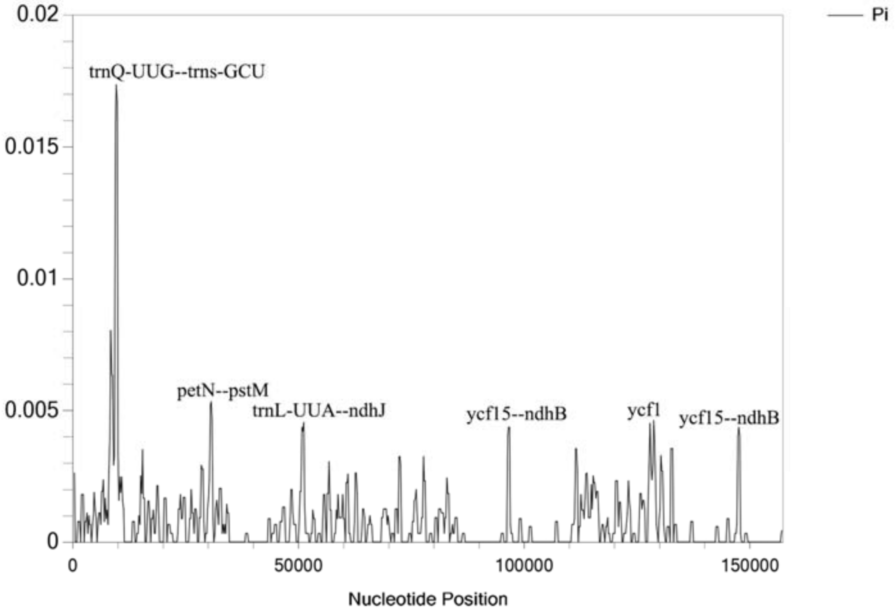
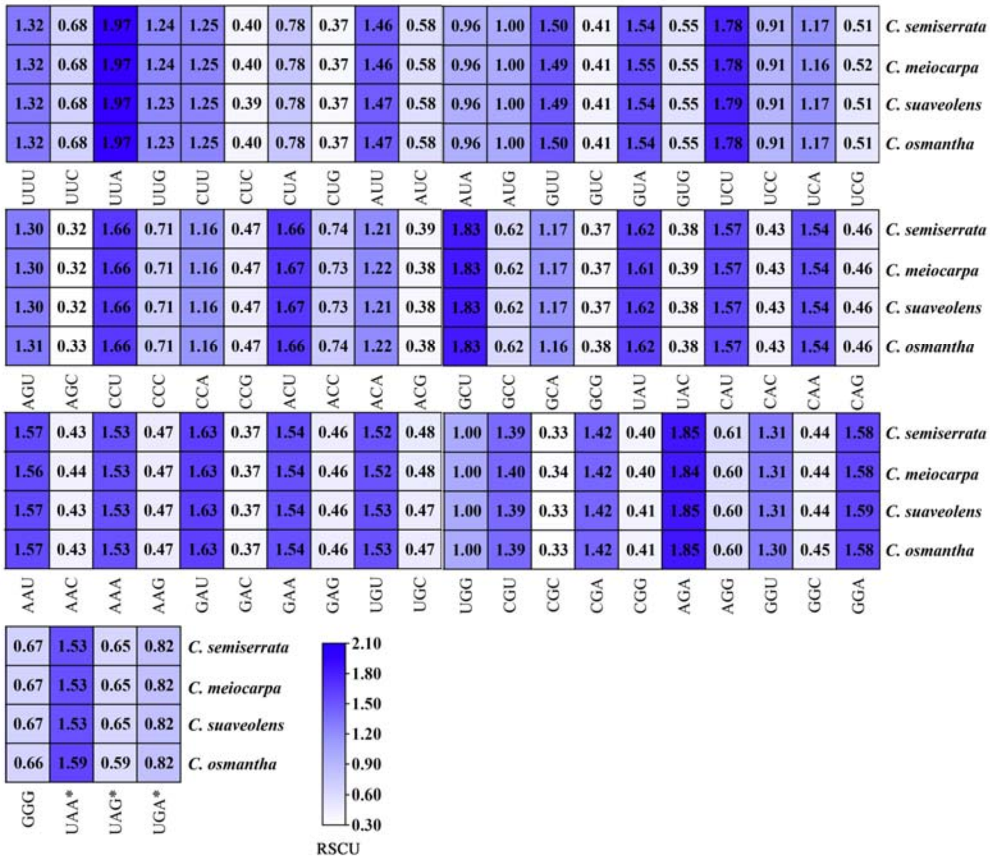


Figure 4. Chloroplast genome sliding window analysis of 10 *Camellia* species. Window length: 2000 bp; step size: 200 bp. X-axis: the position of the midpoint of a window. Y-axis: nucleotide diversity of each window.

Codon Bias Analysis

A total of 88 protein-coding genes (CDS) were identified in the cpDNA of *C. semiserrata*, *C. meiocarpa*, *C. suaveolens*, and *C. osmantha* using Geneious 9.0.2 software. The lengths of the CDS were greater than 300 bp, and the start codon of these CDS was ATG. The stop codons were TAA, TGA, and TAG. Totally of 51 sequences with each kind of stop codon were obtained. There were total of 20,747, 20,743, 20,743, and 20,780 codons in the 51 CDS from *C. semiserrata*, *C. meiocarpa*, *C. suaveolens*, and *C. osmantha*, and the GC content of their CDS was 37.47%, 37.46%, 37.45%, and 37.45%, respectively (**Supplementary Table S2**). RSCU analysis of these codons revealed (**Figure 5** and **Supplementary Table S2**) high conservation in the codon usage of the four *Camellia* plants. The highest coding rate was observed for the codon UUA, which codes for leucine, and UUA comprised 698, 697, 698, and 699 codons in *C. semiserrata*, *C. meiocarpa*, *C. suaveolens*, and *C. osmantha*, respectively. The RSCU value of UUA was 1.97. The lowest coding rate in *C. semiserrata*, *C. meiocarpa*, and *C. suaveolens* was observed for AGC (84), which codes for serine (Ser), and its RSCU value was 0.32. The coding rate of AGC and CGC (85 and 69, respectively), which codes for arginine (Arg), was lowest in C. osmantha. The RSCU value for CGC was 0.33. The RSCU value was greater than one for 29 codons across the four *Camellia* species, with the exception of stop codons. 12 of these codons ended in A, 16 ended in U, and one ended in G. Two codons, AUG and UGG, had RSCU values close to 1, indicating the absence of codon usage bias for these two codons. The RSCU value was less than 1 for 30 codons, with the exception of stop codons. Number of 16 of these codons ended in C, 12 ended in G, and 2 ended in A. These findings indicate that there is a bias for codons ending with A or U in the cpDNA of *C. semiserrata*, *C. meiocarpa*, *C. suaveolens*, and *C. osmantha*.



UUU	UUC	UUA	UUG	CUU	CUC	CUA	CUG	AUU	AUC	AUA	AUG	GUU	GUC	GUA	GUG	UCU	UCC	UCA	UCG	
1.32	0.68	1.97	1.24	1.25	0.40	0.78	0.37	1.46	0.58	0.96	1.00	1.50	0.41	1.54	0.55	1.78	0.91	1.17	0.51	*C. semiserrata*
1.32	0.68	1.97	1.24	1.25	0.40	0.78	0.37	1.46	0.58	0.96	1.00	1.49	0.41	1.55	0.55	1.78	0.91	1.16	0.52	*C. meiocarpa*
1.32	0.68	1.97	1.23	1.25	0.39	0.78	0.37	1.47	0.58	0.96	1.00	1.49	0.41	1.54	0.55	1.79	0.91	1.17	0.51	*C. suaveolens*
1.32	0.68	1.97	1.23	1.25	0.40	0.78	0.37	1.47	0.58	0.96	1.00	1.50	0.41	1.54	0.55	1.78	0.91	1.17	0.51	*C. osmantha*

AGU	AGC	CCU	CCC	CCA	CCG	ACU	ACC	ACA	ACG	GCU	GCC	GCA	GCG	UAU	UAC	CAU	CAC	CAA	CAG	
1.30	0.32	1.66	0.71	1.16	0.47	1.66	0.74	1.21	0.39	1.83	0.62	1.17	0.37	1.62	0.38	1.57	0.43	1.54	0.46	*C. semiserrata*
1.30	0.32	1.66	0.71	1.16	0.47	1.67	0.73	1.22	0.38	1.83	0.62	1.17	0.37	1.61	0.39	1.57	0.43	1.54	0.46	*C. meiocarpa*
1.30	0.32	1.66	0.71	1.16	0.47	1.67	0.73	1.21	0.38	1.83	0.62	1.17	0.37	1.62	0.38	1.57	0.43	1.54	0.46	*C. suaveolens*
1.31	0.33	1.66	0.71	1.16	0.47	1.66	0.74	1.22	0.38	1.83	0.62	1.16	0.38	1.62	0.38	1.57	0.43	1.54	0.46	*C. osmantha*

AAU	AAC	AAA	AAG	GAU	GAC	GAA	GAG	UGU	UGC	UGG	CGU	CGC	CGA	CGG	AGA	AGG	GGU	GGC	GGA	
1.57	0.43	1.53	0.47	1.63	0.37	1.54	0.46	1.52	0.48	1.00	1.39	0.33	1.42	0.40	1.85	0.61	1.31	0.44	1.58	*C. semiserrata*
1.56	0.44	1.53	0.47	1.63	0.37	1.54	0.46	1.52	0.48	1.00	1.40	0.34	1.42	0.40	1.84	0.60	1.31	0.44	1.58	*C. meiocarpa*
1.57	0.43	1.53	0.47	1.63	0.37	1.54	0.46	1.53	0.47	1.00	1.39	0.33	1.42	0.41	1.85	0.60	1.31	0.44	1.59	*C. suaveolens*
1.57	0.43	1.53	0.47	1.63	0.37	1.54	0.46	1.53	0.47	1.00	1.39	0.33	1.42	0.41	1.85	0.60	1.30	0.45	1.58	*C. osmantha*

GGG	UAA*	UAG*	UGA*	
0.67	1.53	0.65	0.82	*C. semiserrata*
0.67	1.53	0.65	0.82	*C. meiocarpa*
0.67	1.53	0.65	0.82	*C. suaveolens*
0.66	1.59	0.59	0.82	*C. osmantha*

Figure 5. RSCU analysis of the cpDNA of four *Camellia* species. Darker colors indicate, greater RSCU and skew. "*" indicates the termination codon.

SSR and Repeat Sequence Analysis

SSR analysis of the cpDNA of four *Camellia* plants and six other plant species was conducted. Dinucleotide and pentanucleotide repeats were not detected in the 10 *Camellia* species (**Figure 6A**). The number of SSRs ranged from 36 to 45, and the greatest number of SSRs was observed in *C. crapnelliana*. The lowest number of SSRs was observed in *C. chekiangoleosa*. Mononucleotide repeats were the most common, followed by tetranucleotide, trinucleotide, and hexanucleotide repeats. The numbers of SSRs in the four *Camellia* species ranged from 38 (in *C. meiocarpa*) to 40 (in *C. osmantha*). Hexanucleotide repeats were not detected in *C. semiserrata* and *C. suaveolens*. The number of mononucleotide and trinucleotide repeats in *C. meiocarpa* was 22 and one, lower than *C. osmantha* , *C. semiserrata* and *C. suaveolens*. The number of SSRs was lowest in *C. meiocarpa* among the four species.

SSRs in the four *Camellia* species were most abundant in the LSC region and least abundant in the IR region (**Figure 6B** and **Figure 6C**). There were two main types of mononucleotide repeats (A/T), and these were only distributed in the LSC and SSC regions. The number of mononucleotide repeats was greater in the LSC region than in the SSC region. Only one type of trinucleotide repeat (TTC) was observed, and it was only present in the LSC. There were 12 types of tetronucleotide repeats across all partitions (AGAT/GTCT/TCTT/TTTC/AAAT/AAAG/TCTA/CCCT/GAAA/ AATA/GAGG/ATAG). The most common tetranucleotide repeat was the AAAT type, and the least common was AAAG. Tetranucleotide repeats were most common in the LSC region, followed by the SSC and IR regions. AAAG was only observed in *C. meiocarpa*. Three hexanucleotide repeats were observed with CTTTTT/AAAAAG/TAAGAT) distributing in the IR region only. These hexanucleotide repeats were most abundant in the gene spacer region and least abundant in introns.

The results of the repeat sequence analysis are shown in **Figure 6D**. The number of repeat sequences in the cpDNA of these 10 *Camellia* species ranged from 37 to 41, and the lengths of these sequences ranged from 30 to 64 excluding the IR region. The palindromic (P) sequences identified were always more abundant than the forward (F) sequences. A total of 37 repeats were detected in *C. semiserrata*, *C. suaveolens*, and *C. osmantha*, including 15 positive repeats and 22 P sequences. There were 39 repeat sequences, 16 F sequences, and 23 P sequences in *C. meiocarpa*.

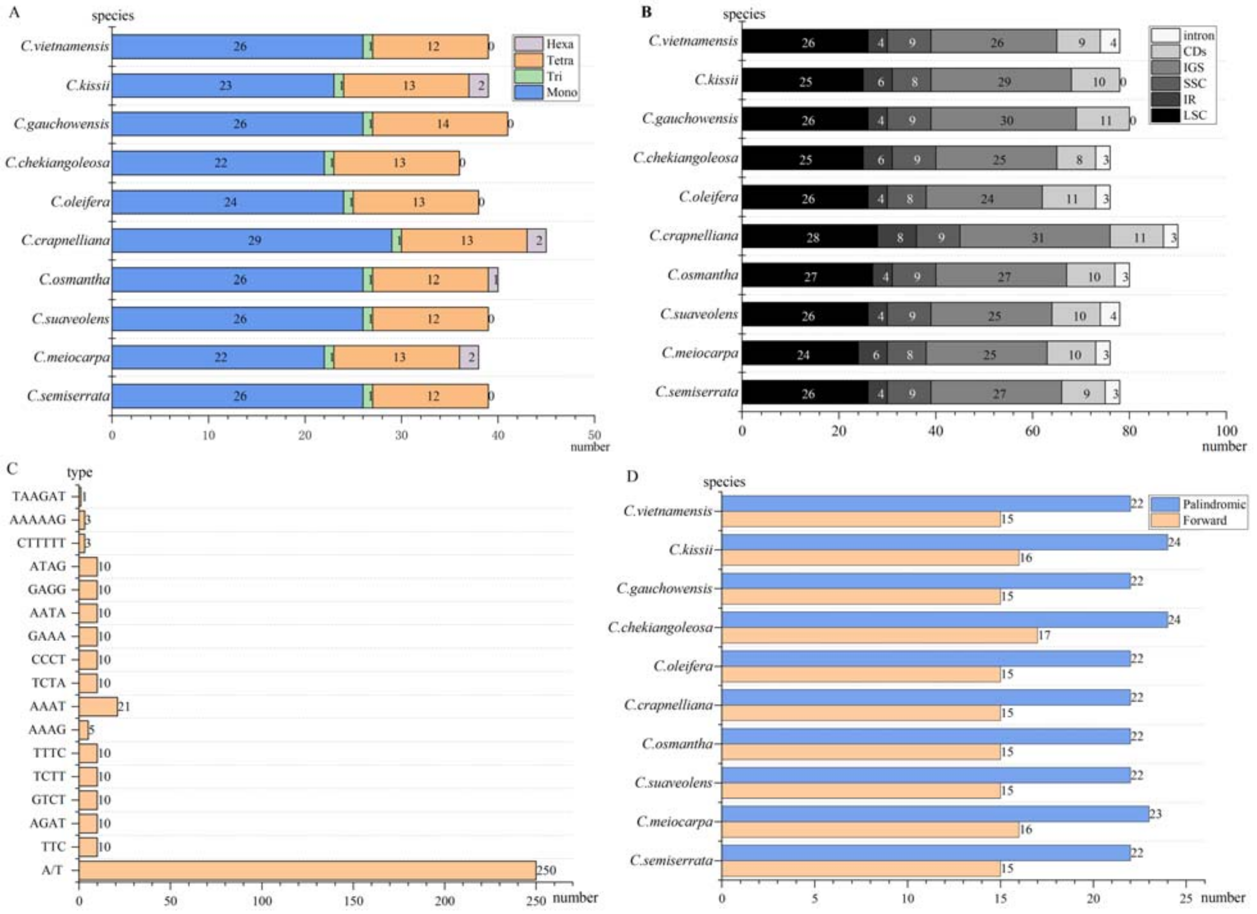


Figure 6. SSRs and INE analysis of the cpDNA of 10 *Camellia* species. X-axis: the number of SSRs or INE; Y-axis: species or SSR type. (A) Number of SSRs; (B) distribution of SSRs; (C) type of SSRs; and (D) type and number of INEs.

Phylogenetic Analysis

The aligned sequences of the complete cpDNA (**Figure 7A**), LSC region (**Figure 7B**), SSC region (**Figure 7C**), IR regions (**Figure 7D**), and shared CDS (**Figure 7E**) of the 35 species were used to construct phylogenetic trees via the Bayesian and ML methods, respectively. Most of the relationships in the phylogenies built based on the complete cpDNA and LSC region had moderate to high support, and the general topology of the trees based on the LSC region and the complete cpDNA was similar (**Figure 7**). In this phylogenetic tree, *C. semiserrata*, *C. osmantha*, and *C. suaveolens* were nested within their own small clades that were clustered into a large clade with *C. meiocarpa*. The phylogenetic trees generated from the first four datasets showed that *C. semiserrata* was most closely related to *C. alzalea*, while *C. suaveolens* was most closely related to *C. gauchowensis*. *C. osmantha* was most closely related to *C. vietnamensis* and was similar to *C. suaveolens* and *C. gauchowensis*. Species of *C. meiocarpa* was most closely related to *C. oleifera*. Phylogenetic trees of *Camellia* species based on the IR regions, the SSC region, and the shared CDS were not highly supported. Thus, the IR regions, the SSC region, and CDS are more conserved and less variable.

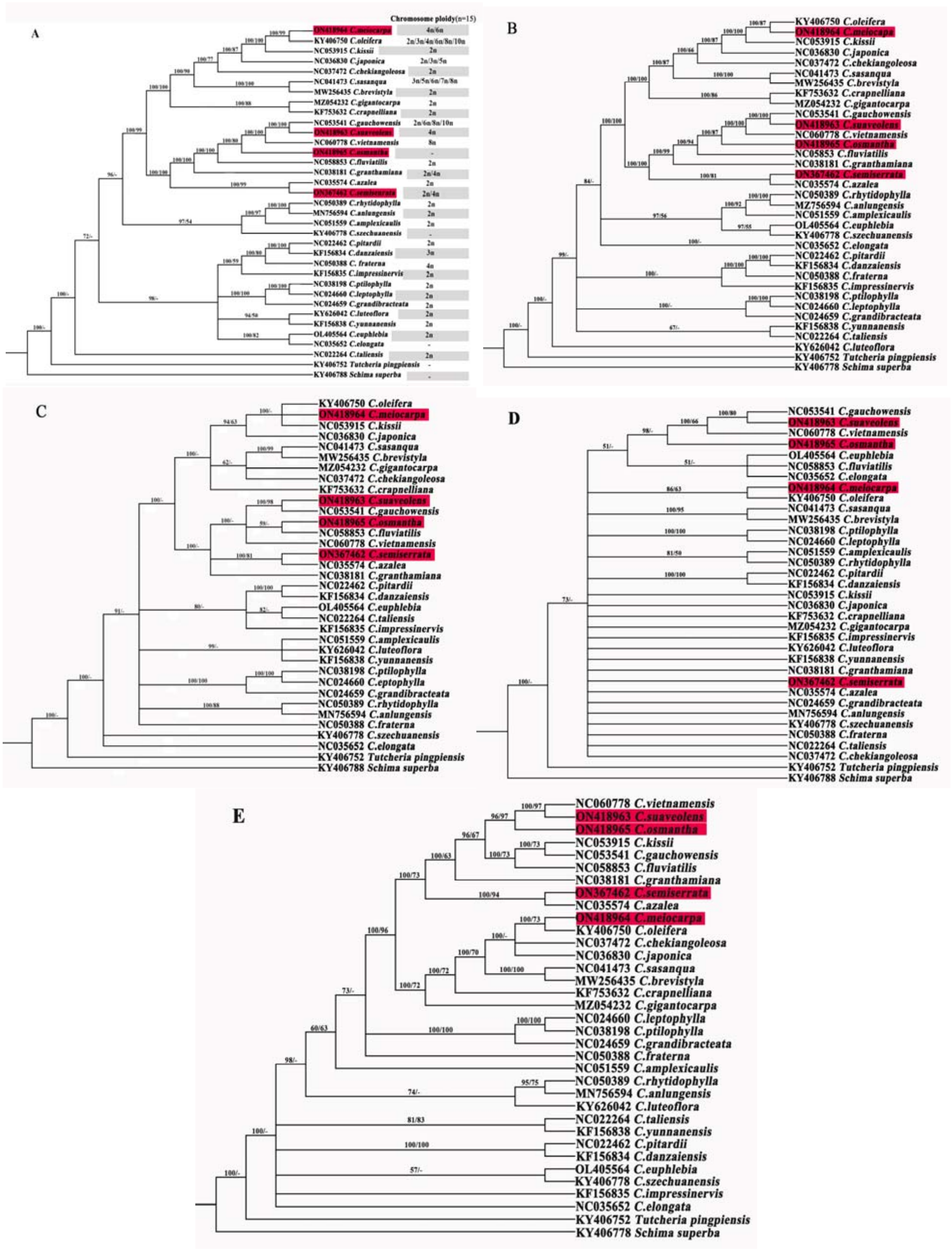


Figure 7. Bayesian phylogenetic tree based on 35 species of complete cpDNA, LSC, SCC, IR and CDS. (A) Bayesian phylogenetic tree based on the complete cpDNA; (B) Bayesian phylogenetic tree based on the LSC region; (C) Bayesian

phylogenetic tree based on SSC regions; (D) Bayesian phylogenetic tree based on the IR regions; and (E) the Bayesian phylogenetic tree based on the CDS. The numbers on the branches are Bayesian and ML bootstrap values.

Discussion

The structure, size, gene content, and genotypes of the cpDNA for most land plants are highly conserved. The lengths of the cpDNA of 100 species in the family Theaceae in the NCBI database range from 150 to 160 kb, and all of them have a typical tetrad structure. The degree of sequence conservation is positively correlated with the GC content (Zheng et al., 2022). The cpDNA of *C. semiserrata*, *C. meiocarpa*, *C. suaveolens*, and *C. osmantha* are highly conserved in their structure, size, gene content, and gene type. The resequenced complete cpDNA of these four *Camellia* species provide new genomic resources for *Camellia* in the NCBI database (Dong et al., 2021; Tong et al., 2022; Chen et al., 2022). The IR region is considered the most conserved region in the cpDNA, and contractions and expansions of the IR boundary play important roles in determining the size and evolutionary trajectory of cpDNA (Wang et al., 2018). In most higher plants, the evolutionary rate of the cpDNA is relatively low, while the sequences and structure are highly conserved. However, the contractions and expansions of IR boundary are commonly observed (Zheng et al., 2022). Contractions and expansions of the IR regions in the four focal *Camellia* species in this study were similar to those observed in *C. weiningensis* (Li et al., 2020), *C. tienii* (Ding et al., 2022), and *C. japonica* (Li et al., 2019). These findings indicate that patterns of IR contractions and expansions vary among *Camellia* species, and this results in changes in the relative lengths of tetrads and full-length cpDNA.

Analysis of nucleotide repeats and molecular makers in cpDNA can be used to distinguish between populations or species, and provide insights into patterns of genetic diversity and systematic relationships (Tang et al., 2022; Guisinger et al., 2010). Highly variable loci in chloroplasts can provide substantial DNA barcoding information that can be used to identify plants. Analysis of the *trnl-trnF*, *rpl16*, and *psbA-tmH* sequences of Sect. Chrysantha in *Camellia* species has shown that *rpl16* and *psbA-tmH* sequences can be used to accurately identify *C. pubipetala* (Chen et al., 2021). The *trnH-psbA* sequence can be used to distinguish plants in different *Camellia* species, but its effectiveness for interspecific classification is weak (Wen et al., 2017). In our study, five highly variable regions were detected (*trnQ-UUG*—*trnG-GCC*, *petN-pstM*, *trnL-UAA*—*ndhJ*, *ycf15*—*ndhB*, and *ycf1*), and these regions could be used in future taxonomic studies of oil-seed *Camellia* species. SSR markers have been widely used to analyze the phylogenetic relationships and genetic diversity of *Camellia* species (Dong et al., 2022; Tao et al., 2019; Wang, 2019; Zhang et al., 2018). Variation was observed in the number of SSRs in the four *Camellia* species and ranged from 36 to 45. The dinucleotide repeats and pentanucleotide repeats were not detected in these four *Camellia* species. No hexanucleotide repeats were detected in *C. semiserrata* and *C. suaveolens*. The numbers of different types of SSRs among the *Camellia* species in this study differ from those of other *Camellia* species in previous studies (Yin et al., 2018). The results of our study showed that SSRs were most common in non-coding regions, followed by regions of coding , LSC , SSC , and IR . The most common mononucleotide repeats were A and T, and no pentanucleotide repeats were observed, which consistent with the results of previous studies (Ding et al., 2022; Yin et al., 2018; Yang et al., 2013). The SSRs of the cpDNA in four *Camellia* species comprised 37 to 39 repeats with lengths ranging from 30 and 64 bp, which is consistent with the results of a previous study of *C. tienii* (Ding et al., 2022). In this study, we did not identify genes or molecular makers in cpDNA that mediated increases in the oil deposited in the seeds as has been identified in other species, such as the long-chain acyl-coA synthetase gene in the FA and TAG synthesis pathways in the leaves of *Suaeda salsa* (Gao et al., 2018).

Codon bias affects mRNA stability, mRNA transcription, and the accuracy of protein translation and protein folding and thus plays a key role in regulating gene expression (Ren et al., 2019). Information on the codon usage of the cpDNA in *Camellia* species, especially comparisons among species, can provide insights into differentially expressed genes, optimize codon usage, and aid the selection of varieties with desirable characters (Teng et al., 2021; Lai et al., 2022; Zhou et al., 2022). The codons of the cpDNA in the four *Camellia* species mainly ended in A/U. This is consistent with the observed codon bias in *C. nitidissima*, *C. oleifera*, and *C. osmantha* in previous studies (Wang et al., 2018; Geng et al., 2022; Hao et al., 2022).

Frequent interspecific hybridization and polyploidy have hindered efforts to resolve the phylogeny and taxonomy of the genus *Camellia*. Comparative analysis of whole cpDNA can provide more reliable insights into phylogenetic relationships among *Camellia* members (Li et al., 2019; Jiang, 2017). Phylogenetic trees built based on the IR dataset, SSC dataset, and CDS were not highly supported, indicating that data from these regions of the cpDNA should not be used in phylogenetic studies of *Camellia* species. The topologies of phylogenetic trees based on the complete cpDNA dataset and the LSC dataset were highly supported (**Figure 7A**). The phylogenetic trees obtained in this study revealed that (1) *C. semiserrata* was most closely related to *C. azalea*; (2) *C. suaveolens* was most closely related to *C. gauchowensis*; and (3) *C. osmantha* was most closely related to *C. vietnamensis*, which was similar to *C. suaveolens* and *C. gauchowensis.*

C. *semiserrata* is more similar to *C. japonica* and *C. chekiangoleosa* based on cytological and morphological characters (Ni, 2007), yet this finding is inconsistent with the results of our study. However, a study of interspecific hybridization of *C. azalea* and *C. semiserrata* has revealed high hybridization affinity among these species (Zhong et al., 2020), and all three of these species are diploid (Jia, 2015). *C. gauchowensis* was considered a synonym of *C. vietnamensis* in Flora of China. However, *C. suaveolens* was more closely related to *C. gauchowensis* than to *C. vietnamensis* according to phylogenetic trees based on cpDNA in this study (**Figure 7A)**. The cross-compatibility between *C. gauchowensis* and *C. suaveolens* via reverse hybridization is low because of their different ploidy. *C. vietnamensis* and *C. gauchowensis* show high cross-compatibility in reciprocal crosses with the same chromosome ploidy (Zhang et al., 2022). The results of ploidy and hybridization analysis revealed that *C. gauchowensis* is more closely related to *C. vietnamensis* than to *C. suaveolens*. *C. suaveolens* has recently been shown to be more similar to *C. furfuracea*, *C. osmantha*, and *C. fluviatilis* based on morphological data and inter-simple sequence repeat markers (Jia, 2015; Wang et al., 2004; Liang et al., 2017). This finding is also not consistent with the results of this study. *C. suaveolens* and *C. gauchowensis* are both decaploid (Zhang et al., 2022). *C. osmantha* shows high affinity with *C. gauchowensis* when hybrids are used as the male parent (Ma et al., 2012). In our study, *C. osmantha* was more closely related to *C. vietnamensis* than to *C. gauchowensis* and *C. suaveolens*. We also found that *C. meiocarpa* and *C. oleifera* were closely related. These findings are consistent with the current classification of these taxa (Jiang, 2017; Ni, 2007). *C. meiocarpa* and *C. oleifera* have been documented to hybridize. *C. meiocarpa* and *C. oleifera* were shown to be genetically closely related via unweighted pair group method with arithmetic mean clustering, and clear gene introgression associated with hybridization between these two species has been observed (Jia, 2015; Zhang et al., 2022; Huang, 2013). According to chromosome ploidy analyses, *C. meiocarpa* is hexaploid (Zhang et al., 2022), and *C. oleifera* ranged in diploid, tetraploid, hexaploid, or octaploid (Huang, 2013; Ye et al., 2020). Polyploidy is the result of interspecific hybridization (Liu and Huang, 2009). The gene exchange associated with interspecific hybridization between *C. meiocarpa* and *C. oleifera* might explain their high genetic similarity (Huang, 2013) and provides support for the taxonomic classification of *C. meiocarpa* and *C. oleifera* proposed by Zhang (1981). Most researchers follow the classification proposed by Hu Xianmu, in which *C. meiocarpa* is considered an independent *Camellia* species (Huang, 2013; Hu, 1957).

Sect. *Oleifera*, Sect. *Camellia*, Sect. *Paracamellia*, and Sect. *Furfuracea* did not form obvious clades in our trees, and many of the branches within these groups were not highly supported. The taxonomic classification of *Camellia* according to cpDNA data differs greatly from that based on morphological data (Lin et al., 2008; Shen et al., 2008; Zhang et al., 2016; Ye, 1988). This indicates that there are limitations associated with phylogenetic studies of *Camellia* based on cpDNA data. First, frequent hybridization and polyploidy of *Camellia* hinder its classification. Second, the cpDNA is uniparentally inherited, and it is evolutionarily conserved. Sequence differences between species are small, and the number of informative loci in the genome is not sufficiently high to permit the resolution of phylogenetic relationships among closely related taxa (Liu et al., 2015).

The cpDNA in this study, along with the results of hybridization and chromosome ploidy analyses, provided new insights into the evolutionary relationships among *Camellia* species, and this phylogeny is more robust than those constructed based on single genes. The general topology of the cpDNA tree is consistent with the classification of *Camellia* based on phenotypic data in this study (Wu et al., 2022). The publication of more nuclear genome sequences and the use of more information from natural hybridization and chromosome structure variation will likely provide further insights into the phylogenetic relationships among *Camellia* species.

Conclusions

In this research, the cpDNA of four *Camellia* species, *C. semiserrata*, *C. meiocarpa*, *C. suaveolens*, and *C. osmantha*, were resequenced, assembled and annotated. We then analyzed the structure of the cpDNA of these four *Camellia* species, as well as contractions and expansions of the IR boundary, nucleotide polymorphism, repeat sequences and SSRs, codon bias, and clarified the phylogenetic relationships within *Camellia* based on hybridization and chromosome information. Our data will aid future studies of the identification, phylogenetic relationships, breeding, and sustainable development of germplasm resources of *Camellia* plants.

Supplementary Material

Supplementary Table S1: Phylogenetic analysis of 35 species；
Supplementary Table S2: RSCU analysis of amino acids in the cpDNA of four *Camellia* species;
Supplementary_data_S1: The aligned sequences of the 35 cpDNA;
Supplementary_data_S2: The aligned sequences of the 35 CDS;
Supplementary_data_S3: The aligned sequences of the 35 IRB region;
Supplementary_data_S4: The aligned sequences of the 35 LSC region;
Supplementary_data_S5: The aligned sequences of the 35 SSC region.

Data Availability Statement

The read sequences of *C. semiserrata*, *C. suaveolens*, *C. meiocarpa* and *C. osmantha* in this study were submitted to the NCBI database under the BioProject ID PRJNA931566 (https://www.ncbi.nlm.nih.gov/sra/PRJNA931566) and BioSample accessions numbers SAMN33062126, SAMN33062127, SAMN33062128, and SAMN33062129. The GenBank accession numbers were as follows: ON367462, ON418963, ON418964, and ON418965.
The name of the plant species *C. osmantha* is referred to as ‘Camellia sp. XJ-2021’ in the NCBI database under the GenBank accession number ON418963. The sequence information for *C osmantha* was published by Ye CX, Ma JL & Ye H. in Nanning, Guangxi, China, in 2012 (Ma et al., 2012). The information for ON418963 was uploaded to the NCBI database.

Acknowledgements

We thank Xiaokeng Forest Farm (Shaoguan City, Guangdong Province, China) for providing plant materials. We thank Guangzhou Nuosai Biotechnology Co., Ltd. for helping us complete chloroplast genome sequencing.

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.

Funding

This research was funded by the Key-Area Research and Development Program of Guangdong Province, grant number 2020B020215003, the Guangzhou Municipal Science and Technology Project, grant number 202201011754, the Forestry Science and Technology Project of Guangdong Province, grant number 2023KJX006, and the Department of Education of Guangdong Province, grant number 2021KQNCX032.

References

State-owned forest farm and forest seedling work station of the State Forestry Administration. **2016.** Oil-tea *Camellia* cultivars in China. Beijing: China Forestry Publishing House. pp1-10.

Li J. X., Huang H. L., Zhao C. Y., He Y. M., Li B., Wu H. 2011. Current situation and breeding direction of *Camellia oleifera* cultivars in Guangdong Province. Guangdong Agricultural Sciences. 20: 34-35 (In Chinese). doi: 10.16768/j.issn.1004-874x.2011.20.072.

Wang J. F., Tan X. J., Wu X. C., Li Q. P., Zhong Q. P., Yan C., Ge X. N. 2020. China's oil tea industry development status and countermeasures. World forestry research. 6: 80-85 doi: 10.13348/j.cnki.sjlyyj.2020.0103.y.

Ma J. L., Ye H., Ye C. X. 2012. *Camellia osmantha*—a new species of *Camellia* Sect.Paracamellia. Journal of Guangxi Botany. 32(06): 753-755. doi: 10.3969/j.issn.1000-3142.2012.06.007.

Liu Y., Xu Y., Jia X. 2021. The complete chloroplast genome of *Camellia osmantha*, an edible oil *Camellia*. Mitochondrial DNA B Resour. 6(11): 3169-3170. doi: 10.1080/23802359. 2021.1987169.

Chang H. T. 1981. A Taxonomy of the Genus *Camellia*. Guangzhou: The editorial staff of the journal of Sun Yat-sen University. 1-183.

Min T. L., Zhang W. J. 1996. Evolution and distribution of *Camellia*. Plant Studies of Yunnan 1. 1-13.

Zhang Y. Z., Zhang K. C., Liao B. Y., Li Y. Q., Zhang M. Y., Li M. Q., Liang R. Y. 2022. Genetic assessment of Hybrid compatibility of *Camellia Oleifera* in Gaozhou based on mixed linear model I. Study on economic forest. 3: 1-13. doi: 10.14067/j.cnki.1003-8981.2022.03.001.

Ye T. W., Yuan D. Y., Li Y. M., Xiao S. X., Gong S. F., Zhang J., Li S. F., Luo J. 2021. Ploidy identification of *Camellia hainanica*. Scientia Silvae Sinicae. 57(7): 61-69. doi: 10.11707 / j.1001-7488.20210707.

Zhong N. Y., Yan D. F., Ke H., Liu X. K., Zhao H. J., Gao J. Y. 2020. Study on interspecific cross-compatibility between *Camellia azealea* and twenty-nine species of *Camellia*. Journal of Anhui Agricultural Sciences. 48(7): 140-145,148. doi: 10.3969/j.issn.0517-6611.2022.07.039.

Chang W. X., Yao X. H., Long W., Ye S. C., Shu Q. L. 2016. Cross-compatibility of four kinds of *Camellia* species. Bulletin of Botanical Research. 36(4): 527-534. doi: 10.7525/j.issn 1673-5102.2016.04.007.

Zhou S., Zhu J. H., Xiao J. Z., Xiang G. H., Shi Y. F. 2001. Experiment on distant hybridization breeding of *Camellia oleifera*. Non-wood Forest Research. 19(1): 20-25. doi: 10.14067/j.cnki.1003-8981.2001.03.007.

Yu Z. Y., Liu Y. J., Xu Y. F., Jia X. C. 2022. Research progress on reproductive biology and hybrid breeding of oil-camelia. Chinese Journal of Tropical Agriculture. 42(11): 44-49. doi: 10.12008/j.issn 1009-2196.2022.11.009.

Tang C. Q., Qiu Z. X., Tan C., Qian Y. M., Chen X. 2022. *Sorbus koenhneana* (Rosaceae): Its complete chloroplast genome and phylogenetic relationship with *S. unguiculata*. Acta Horticulturae Sinica. 3:.641-654 (In Chinese). doi: 10.16420/j.issn.0513-353x.2021-0040.

Tian C. Y., Li Z. Y., Liu Q., Yu L. Q., Wu Z. 2021. Comparison of chloroplast genome structure and phylogenetic analysis of different Alfalfa species. Journal of Chinese grassland. 10: 1-8. dio: 10.16742/j.zGCDXB.20210115.

Zhu B., Qian F., Wang X. S., Liu Y. L. 2022. Phylogeny of Magnoliaceae based on chloroplast genome. Journal of Biology. 3: 53-58. doi: 10.3969/j.issn.2095-1736.2022.03.053.

Rogers, S. O., Bendich A. J. 1989. Extraction of DNA from plant tissues. Berlin Germany: In Plant molecular biology manual. Springer. pp73-83.

Jin J. J., Yu W. B., Yang, J. B., Song Y., de Pamphilis 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(2): 241. doi: 10.1186/s13059-020-02154-5.

Wick R. R., Schultz M. B., Zobel J., Holt K. E. 2015. bandage: interactive visualization of de novo genome assemblies. Bioinformatics. 31(20): 3350-2. doi: 10.1093/bioinformatics/btv383.

Qu X. J., Moore M. J., Li D. Z., 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.

Kearse M., Moir R., Wilson A., Stones-Havas S., Cheung M., Sturrock S., Buxton S., Cooper A., Markowitz S., Duran C., et al. 2021. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 28(12): 1647-1649. doi: 10.1093/bioinformatics/bts199.

Greiner S., Lehwark P., Bock R. 2019. Organellar Genome DRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. Nucl Acid Res. 47(W1): W59-W64. doi: 10.1101/545509.

Chris M., Michael B., Schwartz J. R., Jody S., Alexander P., Edward R., Lior S. P., Inna D. 2000. VISTA: visualizing global DNA sequence alignments of arbitrary length. Bioinformatics. 16 (11): 1046–1047. doi: 10.1093/bioinformatics/16.11.1046.

Amiryousefi A., Hyvönen J., Poczai P. 2018. IRscope: an online program to visualize the junction sites of chloroplast genomes. Bioinformatics. 34(17): 3030–3031. doi:10.1093/bioinformatics/bty220.

Rozas J., Ferrer-Mata A., Sánchez-DelBarrio J. C., Guirao-Rico S., Librado P., Ramos-Onsins S. E., et al. 2017. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Datasets. Molecular Biology and Evolution. 34: 3299-3302. doi: 10.1093/molbev/msx248.

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

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

Sharp P. M., Li W. H. 1987. The codon adaptation index–a measure of directional synonymous codon usage bias, and its potential applications. Nucl Acid Res. 15(3):1281-1295. doi: 10.1093/nar/15.3.1281.

Kazutaka K., John R., Yamada K. D. 2017. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Briefings in Bioinformatics. 4: 1-7. doi: 10.1093/bib/bbx108.

Ronquist F., Teslenko M., van der Mark P., Ayres D. L., Darling A., 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.

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.

Letunic I., Bork P. 2021. Interactive Tree of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Research. 49(W1): W293-W296. doi: 10.1093/nar/gkab301.

Zheng Q., Tong Y. H., Kong Q. B., Feng S. L., Zhou L. J., Ding C. L., Chen T. 2022. Chloroplast genome characteristics and phylogenetic analysis of red Mountain Tea. Journal of Sichuan Agricultural University. 40(04): 574-582 (in Chinese). doi: 10.16036/j.issn.1000-2650.202202074.

Dong L., Yin X., Huang B., Li T., Huang J. J., Wen Q. 2021. The complete chloroplast genome of *Camellia semiserrata* Chi. (Theaceae), an excellent woody edible oil and landscaping species in South China. Mitochondrial DNA Part B. 6(10): 3013-3015. doi: 10.1080/23802359.2021.1976690.

Tong Y. H., Zheng Q., Du X.M., Feng S. L., Zhou L., Ding C. B., Chen T. 2022. Analysis of chloroplast genome sequence of *Camellia polyodonta. J*ournal of Plant Resources and Environment. 31(05): 27-36. doi: 10.3969/j.issn.1674-7895.2022.05.04.

Chen J., Guo Y. J., Hu X. W., Zhou K. B. 2022. Comparison of the chloroplast genome sequences of 13 oil-tea camellia samples and identification of an undetermined Oil-Tea camellia species from Hainan province. Frontiers in Plant Science. 12: 798-581. doi: 10.3389/fpls.2021.798581.

Wang W.C., Chen S.Y., Zhang X. Z. 2018. Whole-genome comparison reveals divergent IR borders and mutation hotspots in chloroplast genomes of herbaceous bamboos (Bambusoideae: Olyreae). Molecules. 23(7): 1537. doi: 10.3390/molecules23071537.

Li Q., Guo Q.Q., Gao Z., Li H. E. 2020. Chloroplast genome characterization of Safflower *Camellia oleifera* from Weining, Guizhou. Acta Horticulturae Sinica. 47(04): 779-787. doi: 10.1007/s12192-020-01117-w.

Ding X. Q., Bi Y.Y., Chen J.T., Xiang S., Lian F.S., Li W. F., Zou S. G. 2022. Analysis of the chloroplast genome of *Camellia tienii*. Jiangsu Agricultural Sciences. 4: 1-8. doi: 10.15889 / j.issn.1002－1302.2022.23.005.

Li W., Zhang C., Guo X., Liu Q., Wang K. 2019. Complete chloroplast genome of *Camellia japonica* genome structures, comparative and phylogenetic analysis. PLoS ONE. 14(5): e0216645. doi: 10.1371/journal.pone.0216645.

Guisinger M. M., Chumley T. W., Kuehl J. V., Boore J. L., Jansen R. K. 2010. Implications of the plastid genome sequence of Typha (Typhaceae, Poales) for understanding genome evolution in Poaceae. Journal of Molecular Evolution. 70(2): 149–166. doi: 10.1007/s00239-009-9317-3.

Chen Y., Guo B. L., Yao L. M., Pan W. G., Zeng J. J., Lu Z. L. 2021. Identification of *Camellia aureus* species based on DNA barcoding. Seed. 40(02): 139-142. doi: 10.1186/s10020-021-00402-3.

Wen B. B., Deng L., Ye X., He W., Liu J. J, Su H. 2017. Application of DNA barcoding in the identification of related *Camellia* species. Guangdong Agricultural Sciences. 44(01): 55-65. doi: 10.16768/j.issn.1004-874x.2017.01.009.

Dond L., Tian Q. Q., Huang B., Xu L. A.,Wen Q. 2022. Association analysis of economic traits with SSR markers in *Camellia chekiangoleosa*. Molecular Plant Breeding. 20(14): 4710-4722. doi: 10.13271/j.mpb.020.004710.

Tao N. Q., Zhang B., Liu X. K., Zhou H. D., Zhong N. S., Yan D. F., Zhang W. J. 2019. Identification of 21 new *Camellia* varieties using SSR markers. Acta Botanica Sinica. 54(01): 37-45. doi: 10.11983/CBB18019.

Wang S. 2019. Based on tea plant transcriptome SSR marker development and part of the plants of the genus *Camellia* genetic diversity analysis. Master's Thesis, Nanjing: Nanjing agricultural university. doi: 10.27244/d.cnki.gnjnu.2019.000006.

Zhang M. H., Zhang H. Q., Cai X. L., Wang H., Ye Y., Wu Y. 2018. SSR analysis of genetic relationship of *Camellia* germplasm resources. Economic Forest Research. 36(04): 130-134. doi: 10.14067/j.cnki.1003-8981.2018.04.020.

Yin X., Wen Q., Wang J. W., Li T., Ye J. S., Xu L. A. 2018. Characterization of microsatellites in complete chloroplast genome of the genus *Camellia* and marker development. Molecular Plant Breeding. 16(20): 6761-6769. doi: 10.13271/j.mpb.016.006761.

Yang J. B., Yang S. X., Li H. T., Yang J., Li D. Z. 2013. Comparative Chloroplast Genomes of *Camellia* Species. PLoS ONE. 8(8): e73053. doi: 10.1371/journal.pone.0073053.

Gao Y. Q., Yan B. W., Zhao Y., Wang F., Dong J. J., He L., Zhao C. J., Li Z. T., Xu J. Y. 2018. Transcriptome analysis of Suaeda salsa and expression profiles of genes related to oil synthesis. Chinese Journal of Oil Crop Sciences. 40(6): 801-811. doi: 10.7505/j.issn.1007-9084. 2018.06. 009.

Ren G. P., Dong Y. Y., Dang Y. K. 2019. Codes in codons: codon bias and fine regulation of gene expression. Science China: Life Sciences. 49(07): 839-847. doi: 10.1360/ssv-2019-0103.

Teng T., Zhao Y. C., Zhao D. G. 2021. Analysis of genetic codon preference of *Camellia sinensis* var. niaowangensis. Genomics and Applied Biology. 40(02): 795-801. doi: 10.13417/j.gab.040.000795.

Zhao Y. M., Yang G. Q., Xu Q. B., Yin X., Xu L., Ding B. 2022. Codon usage bias of chloroplast genome in Acer henryi. Journal of Fujian Agriculture and Forestry university (natural science edition). 51(06): 792-799. doi: 10.13323/j.cnki.j.fafu(nat.sci.).2022.06.011.

Zhou C. Z., Zhu C., Li X. Z., Fu H. F., Lai Z. X., Guo Y. Q. 2020. Research progress of codon usage bias analysis in tea plant. Molecular plant breeding. 18(05): 1480-1488. doi: 10.13271/j.mpb. 018.001480.

Wang P.L., Yang L.P., Wu H. Y., Nong Y. L., Wu S. C., Xiao Y. F., Liu H. L. 2018. Analysis of codon bias in chloroplast genome of *Camellia oleifera*. Journal of Guangxi Botany. 38(02): 135-144. doi: 10.11931/guihaia.gxzw201708001.

Geng X. S., Jia W., Chen J. I., He C. F., Zhu Y. L., Liu Q. 2022. Analysis of codon bias in chloroplast genome of *Camellia aureus*. Molecular plant breeding. 20(07): 2196-2203. doi: 10.13271/j.m pb. 020.002196.

Hao B. Q., Xia Y. Y., Ye H., Gan S. M., Ma J. L. 2022. Analysis of codon bias in chloroplast genome of *Camellia osmantha*. Journal of Central South University of Forestry and Technology. 42(9): 178-186. doi: 10.14067/j.cnki.1673-923x.2022.09.019.

Jiang Z. D. 2017. Based on the chloroplast DNA of plants of the genus *Camellia* molecular systematics and biogeography of preliminary study. Master's Thesis, Zhejiang: Zhejiang university of technology.

Ni S. 2007. Research on phylogenesis Sect. *Camellia* in the genus *Camellia*. Master's Thesis, Najing: Nanjing forestry university.

Zhong N. S, Yan D. F, Ke H., Liu X. K., Zhao H. J., Gao J. Y. 2020. Hybridization affinity between *C. azalea* and 29 native species of *Camellia*. Journal of Anhui Agricultural Sciences. 48(7): 140-145,148. doi: 10.3969/j.issn.0517-6611.2020.07.039.

Jia W. Q. 2015. The *Camellia* reproductive biology and ploidy breeding basic research. Master's Thesis, Beijing: Chinese forestry science institute.

Wang X. J., Shi X. G., Li J. X., Ye C. X. 2004. A new species *Camellia suaveolens* of Sect. *Furfuracea* Chang from Guangdong China. *Journal of Sun Yat-sen University* (Natural Science Edition) 3: 129-130.

Liang G. X., Liu K., Ma J. L., Chen G. C., Ye H., Jiang Z. P. 2017. Molecular classification and identification of *Camellia fragrans*. Economic Forest Research. 35(01): 26-29, 58. doi:10.14067/j.cnki.1003-8981.2017.01.005.

Huang Y. 2013. Population genetic structure and interspecific introgressive hybridization between *Camellia meiocarpa* and *C. oleifera*. Chinese Journal of Applied Ecology. 24(08): 2345-2352. doi: 10.14067/j.cnki.1003-8981.2017.01.005.

Ye T. W., Li Y. M., Zhang Y., Gong Q. Y., Yuan D. Y., Xiao S. X. 2020. Optimization of chromosome section technique and karyotype analysis of *Camellia oleifera*. Journal of Nanjing Forestry University (Natural Science Edition). 44(05): 93-99. doi: 10.3969/j.issn.1000-2006.202003022.

Liu Y. F., Huang H. W. 2009. The dynamics of gene flow and related adaptive evolution in plant populations. Acta Botanica Sinica. 44(03): 351-362 doi: 10.3969/j.issn.1674-3466.2009.03.013.

Hu X. X. 1957. Chinese theaceae I. Science Bulletin. 6: 170.

Lin X. Y., Peng Q. F., Lu H. F., Du Y. Q., Tang B. Y. 2008. Leaf anatomical of *Camellia* Sect. *Oleifera* and Sect.*Paracamellia* (Theaceae) with reference to their taxonomic significance. Journal of Plant Taxonomy. 46(2): 183-193 doi: 10.3724/SP.J.1002.2008.07057.

Shen J. B., Lu H. F., Peng Q. F., Zheng J. F., Tian Y. M. 2008. FTIR spectra of *Camellia* sect. *Oleifera*, Sect. *Paracanmellia*, and Sect. *Camellia* (Theaceae) with reference to their taxonomic significance. Journal of Plant Taxonomy. 46(02): 194-204. doi: 10.3724/SP.J.1002.2008.07125.

Zhang H. Y., Zong X. H., Wang X., Bai X. J., Liang S., Deng H. P. 2016. Population structure and iving community characteristics of endangered *Camellia Luteoflora* Li ex H.T. Plant Science Journal. 34(4): 539-546. doi: 10.11913/PSJ.2095-0837.2016.40539.

Ye C. X. 1988. Taxonomy of *Camellia* and their relationship. Plant Research of Yunnan. 10(1): 61-67.

Liu M., Zhang C. F., Huang J. X., Ma H. 2015. Reconstruction of phylogenetic relationships among Asteraceae (Asteraceae) subfamilies using low copy nuclear genes. Acta Botanica Sinica. 50(05): 549-564. doi: 10.11983/CBB15164.

Wu Q., Tong, W., Zhao H., Ge R., Li R., Huang J., Li F., Wang Y., Mallano A. I., Deng W., Wang W., Wan X., Zhang Z., Xia E. 2022. Comparative transcriptomic analysis unveils the deep phylogeny and secondary metabolite evolution of 116 *Camellia* plants. Plant J. 111: 406-421. doi: 10.1111/tpj.15799.