BMC Plant Biology

RESEARCH

Open Access

Assembly and analysis of the first complete mitochondrial genome sequencing of main Tea-oil Camellia cultivars *Camellia drupifera* (Theaceae): revealed a multi-branch mitochondrial conformation for *Camellia*

Heng Liang[1,3,4,5], Huasha Qi[1,3,4,5], Jiali Chen[1,3,4,5], Yidan Wang[7], Moyang Liu[2], Xiuxiu Sun[1,3,4,5], Chunmei Wang[1,3,4,5], Tengfei Xia[1,3,4,5], Xuejie Feng[3], Shiling Feng[6], Cheng Chen[2] and Daojun Zheng[1,3,4,5*]

Abstract

Background Tea-oil Camellia within the genus *Camellia* is renowned for its premium Camellia oil, often described as "Oriental olive oil". So far, only one partial mitochondrial genomes of Tea-oil Camellia have been published (no main Tea-oil Camellia cultivars), and comparative mitochondrial genomic studies of *Camellia* remain limited.

Results In this study, we first reconstructed the entire mitochondrial genome of *C. drupifera* to gain insights into its genetic structure and evolutionary history. Through our analysis, we observed a characteristic multi-branched configuration in the mitochondrial genomes of *C. drupifera*. A thorough examination of the protein-coding regions (PCGs) across *Camellia* species identified gene losses that occurred during their evolution. Notably, repeat sequences showed a weak correlation between the abundance of simple sequence repeats (SSRs) and genome size of *Camellia*. Additionally, despite of the considerable variations in the sizes of *Camellia* mitochondrial genomes, there was little diversity in GC content and gene composition. The phylogenetic tree derived from mitochondrial data was inconsistent with that generated from chloroplast data.

Conclusions In conclusion, our study provides valuable insights into the molecular characteristics and evolutionary mechanisms of multi-branch mitochondrial structures in *Camellia*. The high-resolution mitogenome of *C. drupifera* enhances our understanding of multi-branch mitogenomes and lays a solid groundwork for future advancements in genomic improvement and germplasm innovation within Tea-oil Camellia.

Keywords Mitochondrial genome, Tea-oil Camellia, Comparative genomics

*Correspondence:
Daojun Zheng
daojunzh@163.com
Full list of author information is available at the end of the article

Background

Tea-oil Camellia (*Camellia* spp.), a member of the genus *Camellia* in the family Theaceae, comprises more than 60 species primarily distributed across southern China and parts of Southeast Asia [1]. Tea-oil Camellia is also an important woody oil species in the world, and *C. drupifera*, *C. meiocarpa* and *C. oleifera* are well-known as the main cultivars of Tea-oil Camellia [2]. Camellia oil, obtained from the mature seeds of these species, is renowned for its unique economic value and high quality, having been used as a cooking oil for over 2,300 years [3]. Notably, Camellia oil is also utilized in a range of domains, including healthcare products, dermatological treatments and the other clinical diagnosis [4].

C. drupifera stands out among Tea-oil Camellia cultivars due to its valuable economic traits, such as high heat resistance, large flowers, and thick skin [1]. It is native to Hainan, Guangzhou, Guangxi in China and Vietnam [5]. *C. drupifera* also exhibits a diversity of ploidy levels, including heptaploid, octaploid, and decaploid variants [2]. Previous studies have highlighted the distinct phenotypes of *C. drupifera* leaves and fruits [6, 7]. Notably, during resource investigations, we observed a unique "one tree with multiple fruits" phenomenon, where the fruits from various branches of the same tree exhibited significant variation in their morphological characteristics. Besides, its seeds contain bioactive compounds such as eriodictyol, taxifolin, and epigallocatechin, which have been linked to pharmacological effects on diseases like cancer, cardiovascular issues, and Alzheimer's disease [8–10]. Moreover, the Camellia oils isolated from *C. drupifera* exhibits the properties of hypoglycemic, hypolipidemic, immunostimulatory, anti-tumor, and anti-inflammatory [11].

In recent years, advancements in molecular biology and genome sequencing technologies have significantly enhanced our understanding of the taxonomy, phylogeny, and population genetics of *C. drupifera* at the molecular level [12–14]. Previous research has indicated that *C. gouchowensis* and *C. vietnamensis* were synonymous species under the unified name *C. drupifera* by the variants in ISSR, SRAP and chloroplast sequence markers [13]. Qi et al. (2023) found that *C. drupifera* from Hainan is an ecotype that is highly differentiated from those in Guangxi and Guangdong [12]. However, all previous studies relied heavily on conventional molecular markers, providing limited genetic insights that may not fully capture the characteristics of the entire genome [15–18]. Based on genome data, clarifying the phylogenetic relationships within *Camellia* would advance taxonomic classification and identification of Tea-oil Camellia [19].

So far, only the complete nuclear genomes of five Tea-oil Camellia species have been published: *C. oleifera* var. "Nanyongensis" [20], *C. lanceoleosa* [21], *C. chekiangoleosa* [22], *C. crapnelliana* [23] and tetraploid *C. oleifera* (main cultivars) [24]. Over 30 complete chloroplast genomes of Tea-oil Camellia species have been published, providing valuable information on the taxonomy and evolutionary relationships within the genus. [25]. However, only one species of Tea-oil Camellia, *C. gigantocarpa*, had its mitochondrial genome sequence available in GenBank (accession number OP270590). Unfortunately, this sequence represents only a fragment of the mitochondrial genome. This gap limits our understanding of mitochondrial genome diversity, evolution, and its potential applications in molecular breeding and species differentiation [26].

It has been suggested that the mitochondria evolved from an ancient endosymbiotic event [27, 28]. Compared to the plant chloroplast genomes, mitochondrial genomes exhibited significant diversity due to lineage-specific evolutionary developments [29, 30]. Many complex structures have been found in plant mitochondrial genomes, including master circular molecules, subgenomic circular forms, linear fragments, and complex branched multigenomic configurations [31]. For instance, the mitochondrial genome of *Panax notoginseng* contains both master circles and subgenomic circles, while recent studies have identified multi-branch structures in other plants [32, 33]. The plant mitochondrial genome was marked by many repetitive sequences and rearrangements, which contributed to its structural diversity. Despite their complexity, mitochondria has preserved a limited set of genes that play crucial roles in regulating oxidative phosphorylation (OXPHOS) and protein translation [34]. This implied that studying mitochondrial genomes will enhance our understanding of the genetic and evolutionary factors which influence mitochondrial evolution. However, the absence of mitochondrial genome data for Tea-oil Camellia species, particularly *C. drupifera*, has limited our ability to fully understand these processes in this economically and ecologically important genus. Even more so, there have been no reports on the comparative analysis of the mitochondrial genomes of *Camellia*.

In this study, we used "3 + 2" method combining Illumina short-read sequencing with Nanopore long-read sequencing to construct the whole mitochondrial genome of *C. drupifera*. Following the assembly and annotation of the mitochondrial genome of *C. drupifera*, the branch structure was found to consist of a larger 900 kb component made up of 18 segments, with the uniting graph being resolved into two linear graphs. The mitochondrial genome of *C. drupifera* was annotated, and its features, phylogenetic relationships, and RNA editing sites were characterized. With these insights, we performed a

comparative examination of *Camellia* species. The findings of this work establish a basis for upcoming genomic investigations and their practical applications in the improvement of Tea-oil Camellia cultivars.

Results

Genome Assembly and Annotation of the Multibranched Mitochondrial Structure in *C. drupifera*

Following sequencing and the removal of nuclear and chloroplast genome-derived sequences, the remaining fragments were assembled to reconstruct the complete mitochondrial genome of *C. drupifera*, using the "3 + 2" method and visualized with Bandage (Fig. 1). The *C. drupifera* mitochondrial genome was characterized by a complex structure and a multi-branched form. The complete mitochondrial genome is 970,986 bp in total length, assembled into 18 segments, with a GC content of 45.73% (Fig. 1). The summary of the assembly statistics were presented in Table 1. In Fig. 1A, the 18 segments range in size from 1,739 (segment 18) to 222,752 bp (segment 1), and in depth from 73.7 (segment 14) to 190.8 (segment 12). Segments 12, 15, 16, and 18, measuring 20,998 bp, 16,475 bp, 11,605 bp, and 1,739 bp respectively, together represented less than 3% of the mitochondrial genome.

Table 1 Summary of assembly statistics

Assembly Statistics	
Number of contigs	18
Largest contig	222, 752
Smallest contig	1,739
Undetermined bases	None
GC content	45.73

Sequencing coverage depth indicates an estimated copy number of two. No sequence variation was evident between the two repeat sequences (labeled as "12" and "12_copy" in Fig. 1B), suggesting that they may be involved in active recombination [35]. For clarity, we organized the genome into two linear representations. Chromosome 1 was structured as a linear sequence of contigs: 10–12-9-16_copy-6–16-11-15_copy-2–18-3–5-13-18_copy-8–1-7–12-17 (Fig. 1C). Chromosome 2 was similarly organized in the order of contigs: 14–15-4. The lengths of Chromosome 1 and Chromosome 2 were 878,626 bp and 92,360 bp, respectively (Fig. 1D), with sequencing depths of 101.6X and 77.6X.

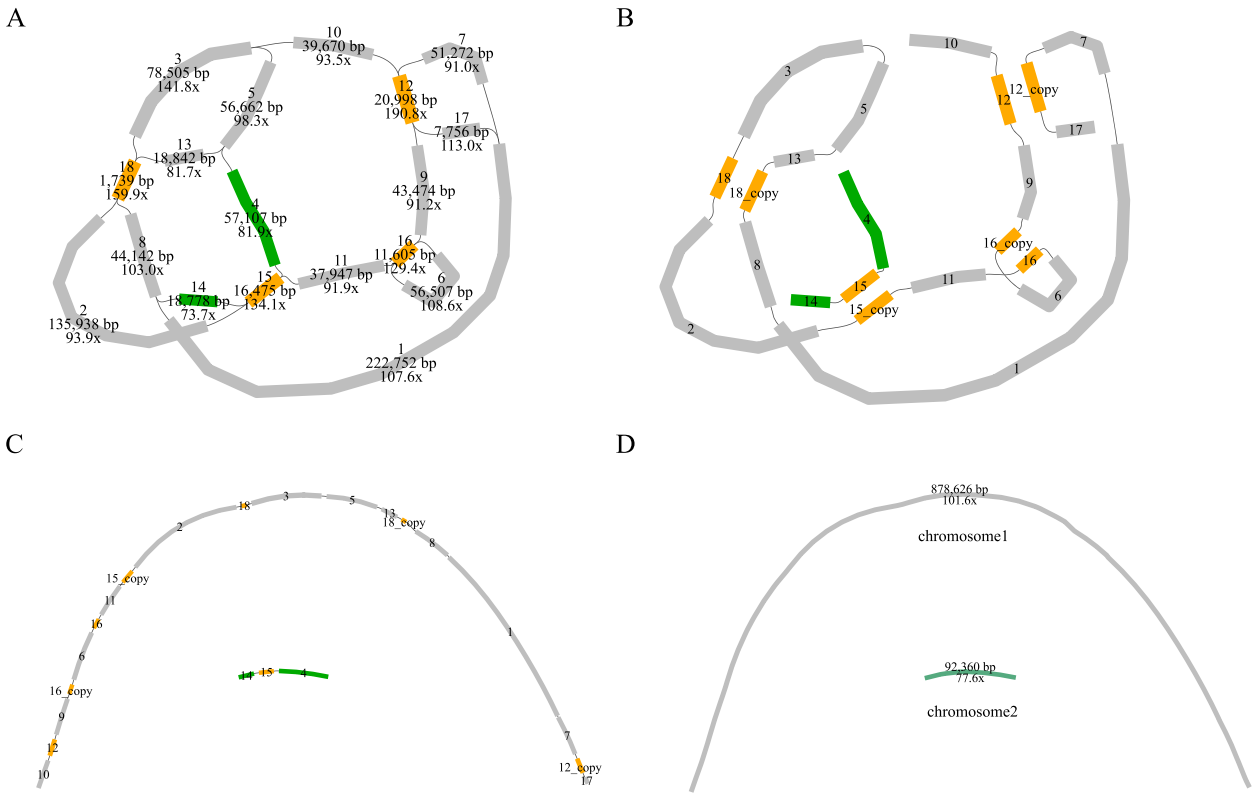

Fig. 1 The assembly result of the mitochondrial genome of *C. drupifera*. **A** the original structure; **B** the simplified structure; **C** re-drawing of **B**; **D** the generated sequence in the end

The *C. drupifera* mitochondrial genome was comprehensively annotated, obtaining 75 genes in total, including 40 protein-coding genes (PCGs), 32 tRNA genes, and three rRNA genes (Table 2). Of the PCGs, 24 were classified as core, while 16 were categorized as non-core. The 24 core genes included five ATP synthase genes (*atp1, atp4, atp6, atp8,* and *atp9*), nine NADH dehydrogenase genes (*nad1*, *nad2*, n*ad3, nad4, nad4L, nad5, nad6, nad7*, and *nad9*), four cytochrome c biogenesis genes (*ccmB, ccmC, ccmFc*, and *ccmFN*), three cytochrome c oxidase genes (*cox1, cox2,* and *cox3*), a transport membrane protein gene (*mttB*), a maturases gene (*matR*), and a Ubichinol cytochrome c reductase gene (*cob*). The non-core genes comprise four ribosomal large subunit genes (*rpl10*, *rpl16*, *rpl2* and *rpl5*), eight small subunits of the ribosome (*rps1*, *rps12*, *rps13*, *rps14*, *rps19*, *rps3*, *rps4* and *rps7*), and two succinate dehydrogenase genes (*sdh3* and *sdh4*). The relative arrangement and orientation of these genes are shown in Fig. 2.

Mitochondrial Genomic Comparison between *C. drupifera* and Other *Camellia* Species

To investigate the evolutionary dynamics of the *C. drupifera* mitochondrial genome, we compared it with six other *Camellia* species. The GC content in these genomes varied from 45.49% (*C. gigantocarpa*) to 45.75% (*C. sinensis*). The genome sizes of the six *Camellia* species ranged from 707,441 bp (*C. sinensis*) to 1,081,966 bp (*C. sinensis* var. *assamica*). Moreover, the number of rRNAs, tRNAs, introns and PCGs varied, ranging from 2 to 4 rRNAs, 18 to 32 tRNAs, 7 to 33 introns, and 32 to 47 PCGs, respectively (Table 3).

The mitochondrial genomes of these *Camellia* species exhibited minimal variation in GC content, while the

Table 2 The protein-coding genes of the mitochondrial genome of *C. drupifera*

Group of genes	Gene name
ATP synthase	*atp1 atp4 atp6 atp8 atp9*
Cytohrome c biogenesis	*ccmB ccmC ccmFc* ccmFn*
Ubichinol cytochrome c reductase	*cob*
Cytochrome c oxidase	*#cox2 cox1 cox2 cox3*
Maturases	*matR*
Transport membrance protein	*mttB*
NADH dehydrogenase	*nad1**** nad2****(2) nad3 nad4** nad4L nad5**** nad6 nad7**** nad9*
Ribosomal proteins (LSU)	*rpl10 rpl16 rpl2* rpl5*
Ribosomal proteins (SSU)	*#rps19 #rps7 rps1 rps12(2) rps13 rps14 rps19 rps3* rps4 rps7*
Succinate dehydrogenase	*sdh3 sdh4*
Ribosomal RNAs	*rrn18 rrn26 rrn5*
Transfer RNAs	*trnA-TGC* trnC-GCA(2) trnD-GTC trnE-TTC trnF-AAA* trnF-GAA trnG-GCC trnH-GTG trnI-GAT*(2) trnK-TTT trnM-CAT(6) trnN-ATT trnN-GTT trnP-TGG trnQ-TTG trnS-GCT(2) trnS-TGA trnS-TGA* trnT-GGT* trnT-TGT* trnV-GAC trnW-CCA(2) trnY-GTA*
Other	

Gene*: intron number; # Gene: Pseudo gene; *Gene(2)*: Number of copies of multi-copy genes

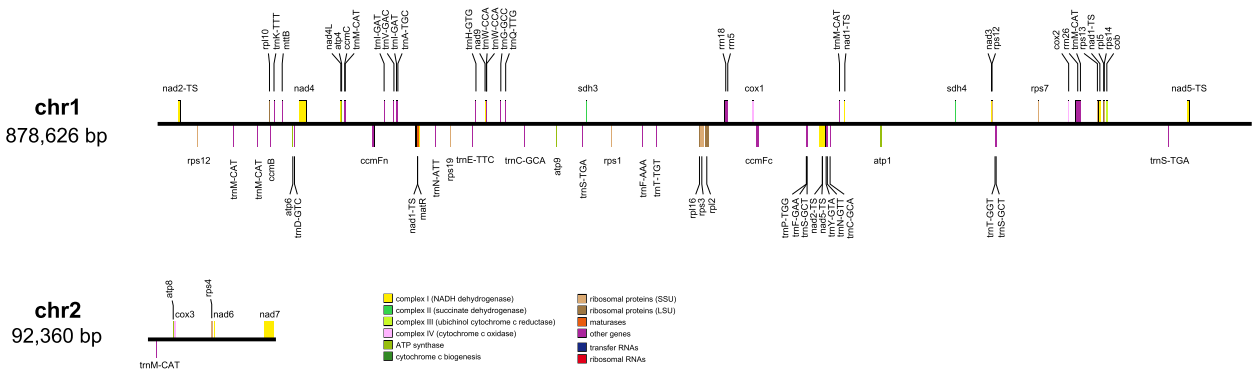

Fig. 2 The order, orientation, and size of the genes within the *C. drupifera* mitochondrial genome

gene count varied considerably. According to our results, *C. sinensis* var. *assamica* (GenBank: OL989850) had the highest gene count, while *C. gigantocarpa* had the fewest. The reason is probably that the mitochondrial genomes of *C. gigantocarpa* are incomplete. Compared to the six other *Camellia* species, *C. drupifera* ranks as the second-largest in terms of gene number, GC content, and size (second only to *C. sinensis* var. *assamica*, GenBank: OL989850).

To evaluate the variation in PCGs and clarify evolutionary patterns, we calculated the non-synonymous/synonymous substitution (Ka/Ks) and nucleotide diversity (Pi) for PCGs across all *Camellia* species. The average Pi for individual genes ranged from 0 to 0.08806 (Table S1). In Fig. 3A, in comparison to *C. drupifera*, only one gene (*rpl2*) in *C. gigantocarpa* exhibits a Ka/Ks ratio above 1, suggesting it has undergone positive selection during evolution. Other genes (like *atp1*, *ccmFc* and *nad2* et al.) were going through purify selection. Additionally, the Pi value for the gene *rrn18* was the highest at 0.08806, while the Pi values for 14 other genes were 0, indicating no observed nucleotide diversity. Consistent genetic distance patterns were observed among PCGs, with *rrn18* (0.08806), cox2 (0.01885), *atp9* (0.01714), *nad5* (0.01197), and *ccmFc* (0.00763) identified as fast-evolving genes then other PCGs. In contrast, *cox1* (0.00027), *rps3* (0.00051), and *rrn26* (0.00054) were noted as slow-evolving genes (Fig. 3B).

SSRs and tandem repeats of *C. drupifera* mitochondrial genome

In plant mitochondrial genomes, repetitive sequences are essential for their evolutionary development [36]. Simple sequence repeats (SSRs) are short motifs, typically 1 to 6 bp in length, arranged in tandem [37]. A total of 269 SSRs were detected in the mitochondrial genome of *C. drupifera* (Fig. 4 and Table S2). Among the repeat sequences, the most abundant repeat sequences were tetra-nucleotides, making up 108 loci (40.15%). Followed by di-nucleotide repeats with 73 loci (27.14%), tri-nucleotide repeats with 40 loci (14.87%), mono-nucleotide repeats with 27 loci (10.04%), penta-nucleotide repeats with 17 loci (6.32%), and hexa-nucleotide repeats with 4 loci (1.48%). In *Camellia* species, a total of 201 to 316 SSRs were found (Fig. 5 and Table S2). Tetra-nucleotide repeats were the most common, whereas hexa-nucleotide repeats were the least frequent (Table S2). The total number of SSRs showed a weak correlation with mitochondrial genome sizes (Table S2), implying that the increase in repeat sequences may not significantly contribute to genome enlargement in *Camellia*.

Tandem repeats, which consist of two or more consecutive copies of a nucleotide pattern, emerge through the duplication of adjacent genomic regions [38]. *C. drupifera*'s mitochondrial genome contains 43 tandem repeats, with lengths spanning from 5 to 61 bp (Fig. 4 and Table S3). Dispersed repeats are another type of repeat sequences, differing from tandem repeats in their organizational form [39]. Dispersed repeats are scattered throughout the genome, often existing as moderately repetitive sequences. In *C. drupifera*, 802 dispersed repeats were observed, with lengths extending from 29 to 21,502 bp (Fig. 4 and Table S4).

Analysis of codon usage in *C. drupifera* mitochondrial genome

In *C. drupifera* mitochondrial genome, a total of 10,500 codons were found (Table 4). The mitochondrial DNA of *C. drupifera* encoded all 20 standard amino acids, and 61 distinct codon types were observed. The most frequently occurring codon was UAA, a stop codon (Table 4). Leucine was found to be the most frequently encoded amino acid, with 1,077 codons (10.26% of the total), followed by Serine with 975 codons (9.29%). On the other hand, Cysteine had the fewest codons, totaling only 149 (1.42%). We identified 31 codons that appeared more frequently than expected (RSCU > 1) and other were RSCU < 1. Tryptophan (UGG) and Methionine (AUG) showed no codon preference, both having an RSCU value of 1. Excluding these two, most

Table 3 General characteristics of six *Camellia* mtDNAs

	C. drupifera	*C. sinensis*	*C. sinensis*	*C. sinensis* var. *assamica*	*C. sinensis* var. *assamica*	*C. nitidissima*	*C. gigantocarpa*
Genbank	PQ041261-PQ041262	MH376284	OM809792	MK574876	OL989850	ON645224	OP270590
Size(bp)	971,986	707,441	914,855	880,048	1,081,966	949,915	970,410
GC%	45.68	45.75	45.66	45.57	45.62	45.71	45.49
rRNAs	3	2	3	3	4	3	3
tRNAs	32	23	30	23	30	29	18
introns	28	14	25	18	33	15	7
PCGs	39	32	42	40	47	36	38

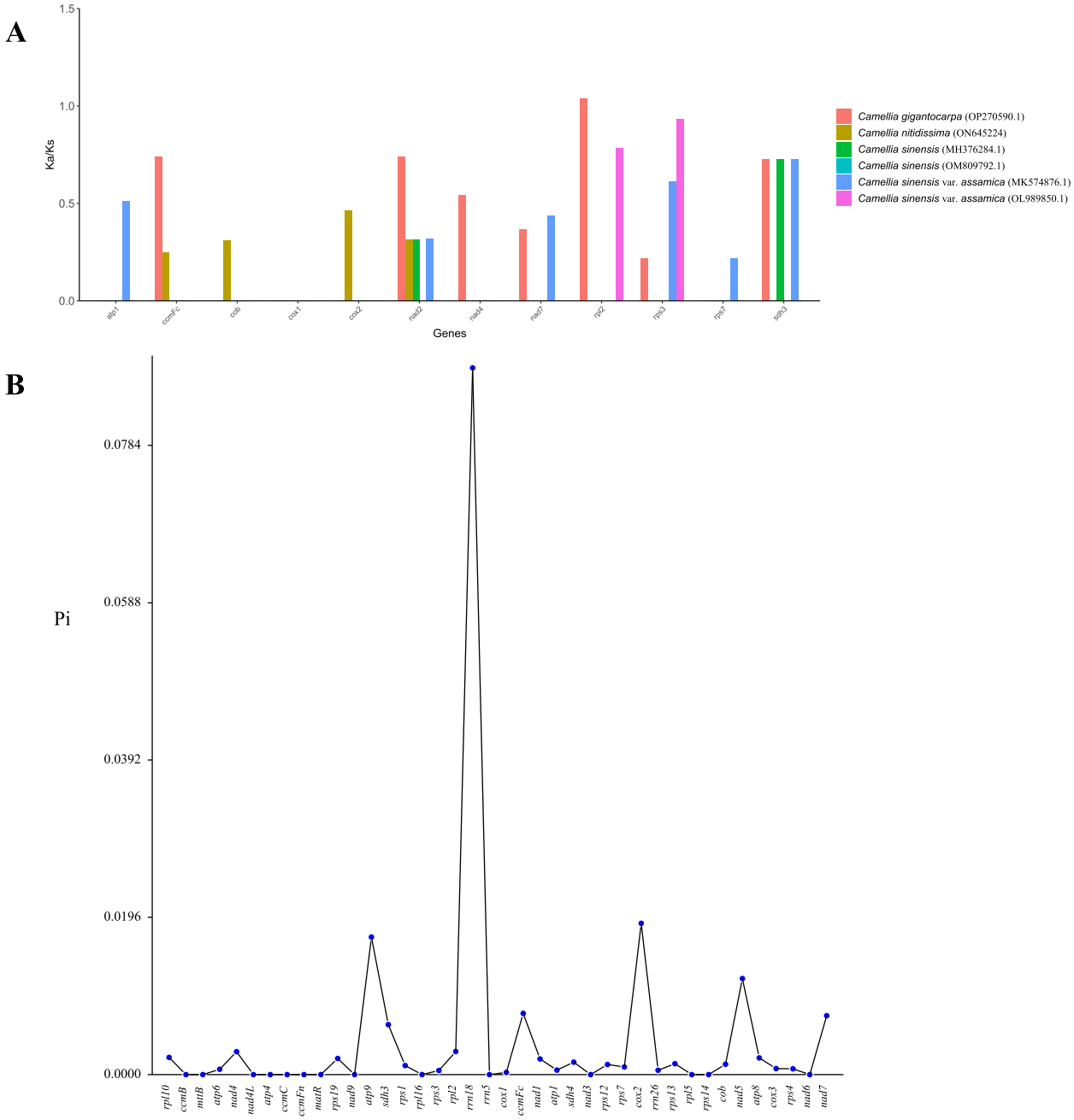

Fig. 3 Variation in mitochondrial genes and the evolutionary characteristics of *Camellia*. **A** Ka/Ks ratio calculated for the PCGs. **B** nucleotide diversity (Pi) of the PCGs

amino acids displayed significant bias in codon usage (Fig. 6). Amino acids like Arginine, Leucine, and Serine are encoded by multiple codons, with each having six possible codons.

To further investigate codon usage bias in *C. drupifera*, we extracted PCGs (*nad2, rps12, rpl10, ccmB, mttB, atp6, nad4, nad4L, atp4, ccmC, ccmFn, nad1, matR, rps19, nad9* and *atp9* etc.) from the *C. drupifera* mitochondrial DNA (Table S5). The GC content of the first (GC1), second (GC2), and third (GC3) positions of these genes were calculated, and the results indicated that the values spanned from 36.88% to 57.84% for GC1, from 35.51% to 55.86% for GC2, and from 23.93% to 58.38% for GC3. At different positions, the GC content varied between 37.32% and 52.39%, reflecting a bias towards A/T base pairs and A/T-terminated

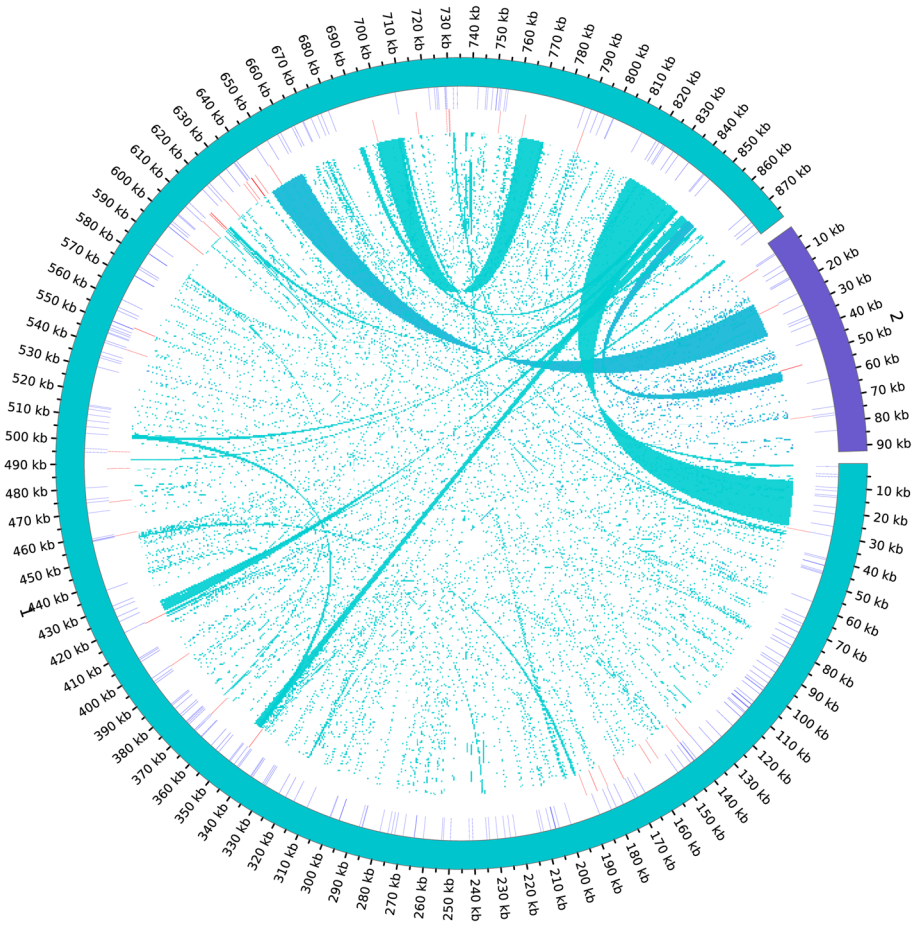

Fig. 4 Distribution of repetitive sequences in *C. drupifera* mitochondrial genome. The outermost circle represents the mitochondrial genome; the inner circles are SSR (Blue), tandem repeat (red), and dispersed repeat (turquoise)

codons in *C. drupifera*. Additionally, we computed the effective number of codons (ENC) for these protein-coding genes, spanning from 33.62 to 61. The average ENC exceeded 35, indicating a relatively weak codon usage bias. Furthermore, in a neutrality plot analysis of *C. drupifera* mitochondrial DNA, a correlation of 0.143 was observed between GC12 and GC3, with a significance level ($P = 0.05$) lower than anticipated (Fig. 7A). This result indicates that codon usage bias in *C. drupifera*'s mitochondrial DNA is largely influenced by natural selection. To better understand the determinants of codon usage in *Camellia*, the ENC values were calculated and plotted against GC3 values (Fig. 7B). The ENC-plot indicated that most of genes were positioned below the standard curve, with only a few above it. This suggests that the selection pressure influenced codon preferences in the mitochondrial genome of *C. drupifera*.

Chloroplast-to-mitochondrial gene transfer in *C. drupifera*

During the evolution of higher plants, genetic material is frequently transferred between cellular organelles, particularly within mitochondrial and chloroplast genomes [40]. However, chloroplast-derived sequence fragments tend to demonstrate relatively lower conservation [41]. To explore this phenomenon in *C. drupifera*, we conducted a sequence similarity analysis aimed at identifying instances of sequence migration from the chloroplast to the mitochondrion (Fig. 8). We identified over 20 homologous fragments shared between *C. drupifera* chloroplast and mitochondrial genomes. These fragments ranged in alignment lengths from 15 to 505 bp, with mismatches ranging from 0 to 212. The total length of these fragments is 16,785 bp, representing 1.73% of the mitochondrial DNA (27,468 bp) and 17.5% of the chloroplast DNA in C. *drupifera*, and these fragments are referred to as MTPTs (Fig. 8). Upon annotating these sequences, we identified seven complete tRNA genes (*trnV-GAC; trnI-GAT; trnA-TGC; trnM-CAT; trnN-GTT; trnD-GTC* and *trnW-CCA*).

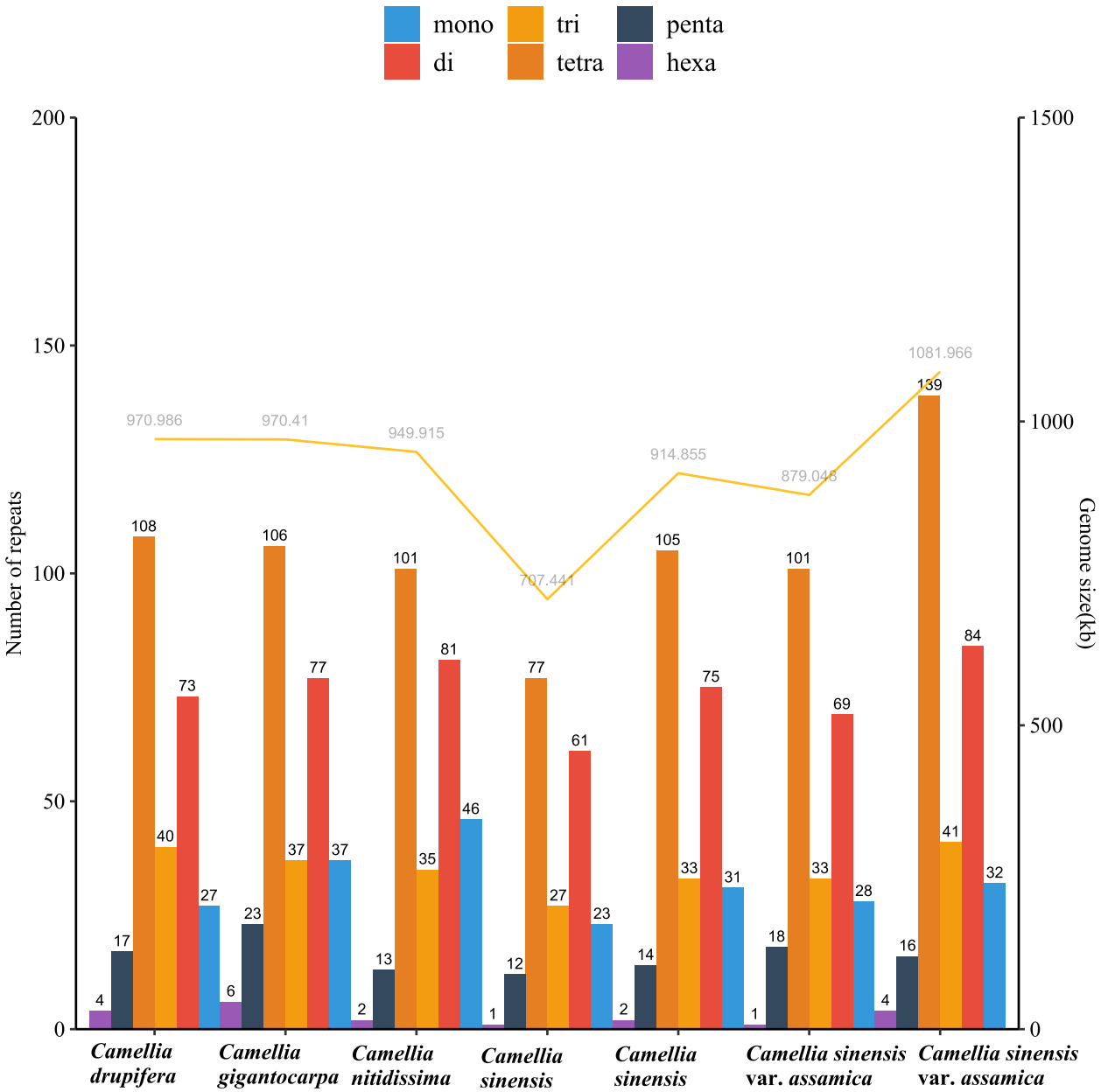

Fig. 5 The SSRs in *Camellia* species

Analysis of collinearity among *C. drupifera* mitochondrial genome compared with other *Camellia* species

BLASTN was used for comparative analysis of the mitochondrial genome of *C. drupifera* with other *Camellia* species, allowing us to identify homologous genes and their sequence arrangement. We focused on conserved collinearity blocks of 500 bp or more, and blocks longer than 0.5 kb were retained for further analysis to enhance the visualization of collinearity patterns (Fig. 9). This analysis revealed numerous homologous collinear blocks, though they tended to be relatively short in length. Importantly, conserved genes (including *atp8, atp9, ccmB, ccmC, ccmFc, ccmFn, cob, cox2, matR, mttB, nad1, nad2, nad3, nad4, nad4L, nad5, nad6, rpl10, rpl5, rps1, rps12, rps12-2, rps13, rps14, rps19, rps4, rrn26, sdh3, trnC-GCA, trnD-GTC, trnF-GAA, trnK-TTT, trnM-CAT-2, trnM-CAT-3, trnM-CAT-4, trnM-CAT-5, trnM-CAT-6, trnN-GTT, trnP-TGG, trnS-GCT, trnS-GCT-2,* and *trnY-GTA*), were identified in the homologous

Table 4 Relative synonymous codon usage in *C. drupifera* mitochondrial genome

Symbol	Codon	No	RSCU	Symbol	Codon	No	RSCU
Ter	UAA	20	1.5789	Met	AUG	275	1
Ter	UAG	6	0.4737	Asn	AAC	113	0.6828
Ter	UGA	12	0.9474	Asn	AAU	218	1.3172
Ala	GCA	168	0.9912	Pro	CCA	174	1.1658
Ala	GCC	165	0.9735	Pro	CCC	111	0.7437
Ala	GCG	86	0.5074	Pro	CCG	97	0.6499
Ala	GCU	259	1.528	Pro	CCU	215	1.4405
Cys	UGC	53	0.7114	Gln	CAA	221	1.5137
Cys	UGU	96	1.2886	Gln	CAG	71	0.4863
Asp	GAC	99	0.6018	Arg	AGA	177	1.4351
Asp	GAU	230	1.3982	Arg	AGG	94	0.7622
Glu	GAA	294	1.3425	Arg	CGA	156	1.2649
Glu	GAG	144	0.6575	Arg	CGC	75	0.6081
Phe	UUC	294	0.9145	Arg	CGG	87	0.7054
Phe	UUU	349	1.0855	Arg	CGU	151	1.2243
Gly	GGA	271	1.4688	Ser	AGC	97	0.5969
Gly	GGC	100	0.542	Ser	AGU	165	1.0154
Gly	GGG	131	0.71	Ser	UCA	188	1.1569
Gly	GGU	236	1.2791	Ser	UCC	159	0.9785
His	CAC	60	0.4563	Ser	UCG	140	0.8615
His	CAU	203	1.5437	Ser	UCU	226	1.3908
Ile	AUA	225	0.8142	Thr	ACA	131	0.9668
Ile	AUC	240	0.8685	Thr	ACC	142	1.048
Ile	AUU	364	1.3172	Thr	ACG	81	0.5978
Lys	AAA	268	1.1703	Thr	ACU	188	1.3875
Lys	AAG	190	0.8297	Val	GUA	189	1.185
Leu	CUA	159	0.8858	Val	GUC	112	0.7022
Leu	CUC	110	0.6128	Val	GUG	141	0.884
Leu	CUG	99	0.5515	Val	GUU	196	1.2288
Leu	CUU	235	1.3092	Trp	UGG	152	1
Leu	UUA	257	1.4318	Tyr	UAC	73	0.4591
Leu	UUG	217	1.2089	Tyr	UAU	245	1.5409

collinear blocks of *C. drupifera* and other *Camellia* species, showing over 99% sequence identity.

Predict RNA editing sites

We predicted a total of 531 RNA editing sites within 38 protein-coding genes (PCGs) of the *C. drupifera* mitochondrial genome (Fig. 10). Among these, the *ccmFn* gene had the highest number of editing sites, with 40 identified, followed by the *ccmB* gene, which had 35. In addition, the *rps1, rpl10, rps14, rps19, rps7, sdh3 and sdh4* genes each had two or three RNA editing events, which were associated with the function of ribosomal proteins and succinate dehydrogenase. The first and second codon positions were the main sites of RNA editing-induced amino acid modifications, with the second position being the most frequently altered. [42]. Our results are consistent with previous findings, such as Arginine (R) to Tryptophan (W), Alanine (A) to Valine (V), and Serine (S) to Leucine (L), which play an important role in increasing protein stability (Table S6).

Phylogenetic analysis

Mitochondrial PCG nucleotide sequences were extracted from a selection of species, including seven *Camellia*, two *Solanum*, one *Nicotiana*, one *Helianthus*, one *Platycodon* species and two outgroup species *Arabidopsis thaliana* and *Brassica rapa*. The phylogenetic trees were generated using both ML and BI approaches (Fig. 11). Of the 13 nodes on the phylogenetic tree, eight exhibited bootstrap support values

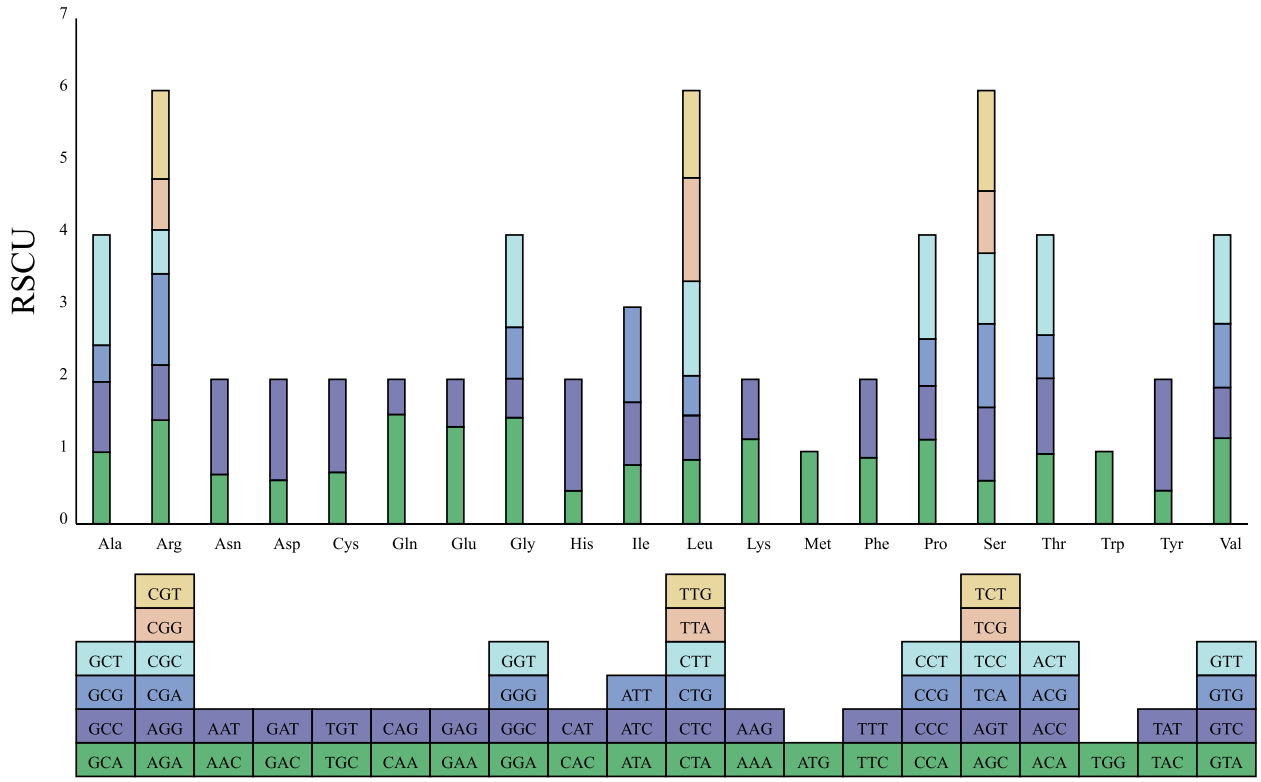

Fig. 6 Analysis of RSCU in *C. drupifera* mitochondrial genome

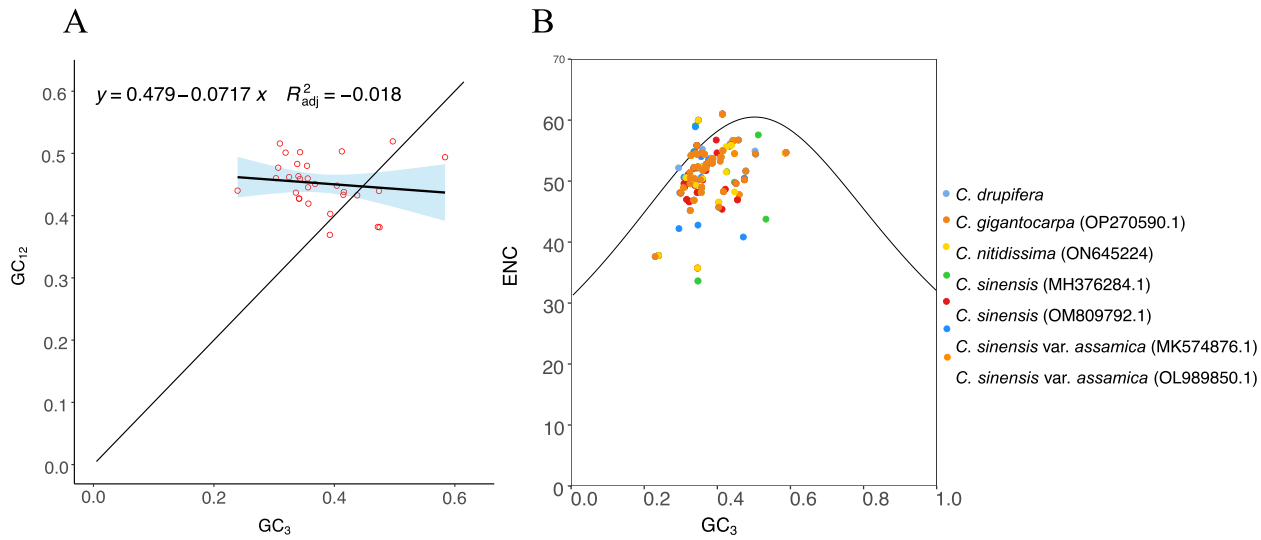

Fig. 7 **A** GC content of different positions from PCGs. **B** ENC-plot against GC3 of mitochondrial genome of *C. drupifera*

greater than 80%, with posterior probabilities of 1.0. Notably, only 2 out of 7 nodes were well-supported for the Theaceae clade (Fig. 11A). Among the eight Theaceae species, *C. drupifera* formed a basal clade (BS = 100, PP = 1.0). *Stewartia sinensis* clustered with other *Camellia* species but with low support (BS < 80, PP < 1.0).

To explore the evolutionary connections between mitochondrial and chloroplast genes, we accessed the chloroplast genome sequences for the same species

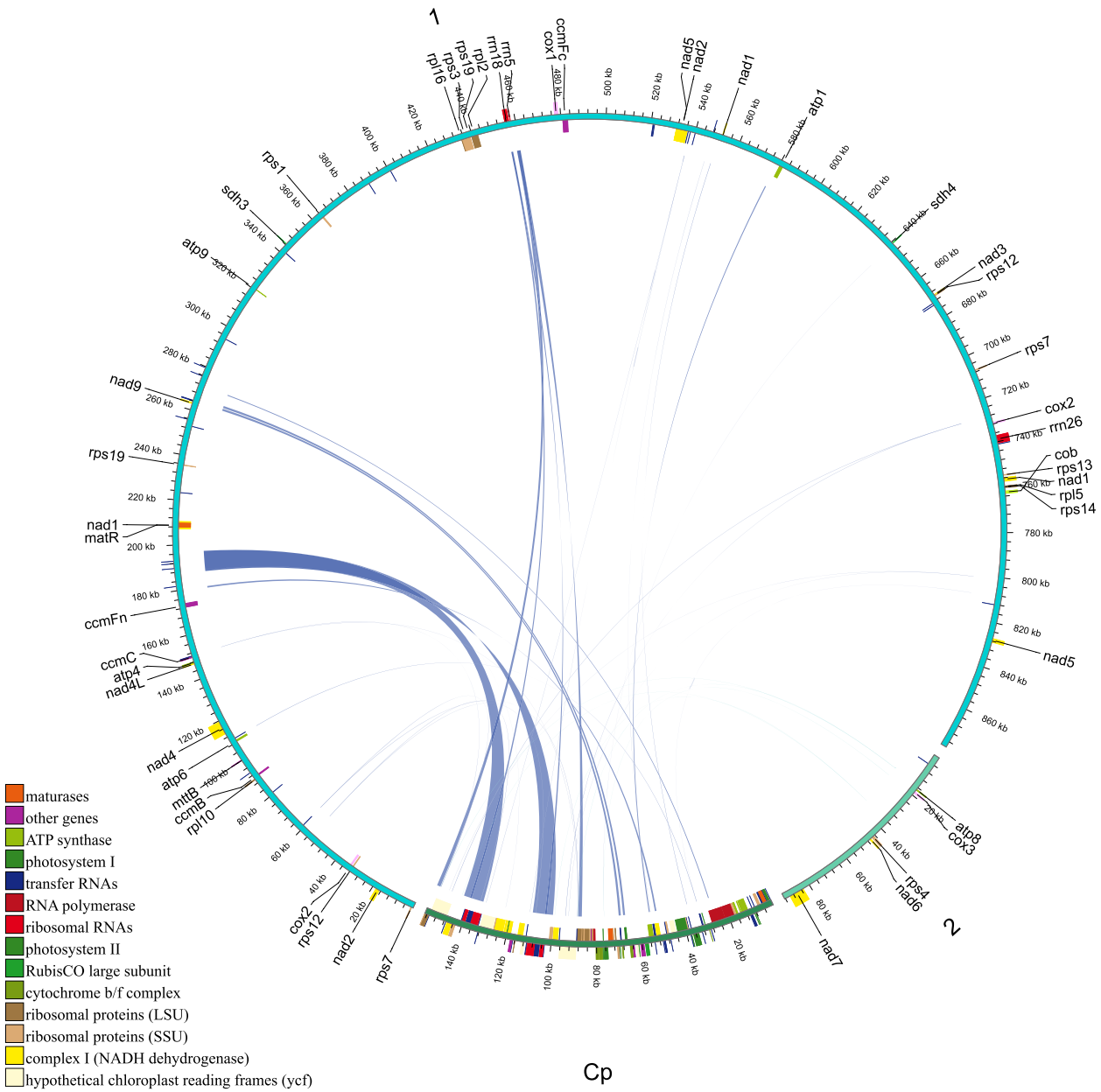

Fig. 8 Homologous analysis based on different organelles shows the cyan arc representing mtDNA and the green arc representing the chloroplast genome. Yellow lines between the blue arcs indicate homologous fragments

from GenBank. Phylogenetic analysis of conserved PCG sequences were conducted using the same methods applied in the mitochondrial genome analysis (Fig. 11B). Our analysis produced a phylogenetic tree with a more reliable topological arrangement, which was different with the one derived from mitochondrial PCGs. 12 in 13 nodes on the phylogenetic tree exhibited a higher support (BS > 80 and PP = 1.0) than mitochondrial dataset. Among the Theaceae species, *Stewartia sinensis* was basal clade (BS = 100 and PP = 1.0), which was regarded as an early-diverging genus in Theaceae [19]. The *Camellia* species clustered into a single branch (BS = 100, PP = 1.0), with the two Tea-oil Camellia species, *C. drupifera* and *C. gigantocarpa*, forming a subclade. In contrast, in the mitochondrial-derived tree, *C. nitidissima* was clustered together with *C. gigantocarpa* and *C. sinensis*.

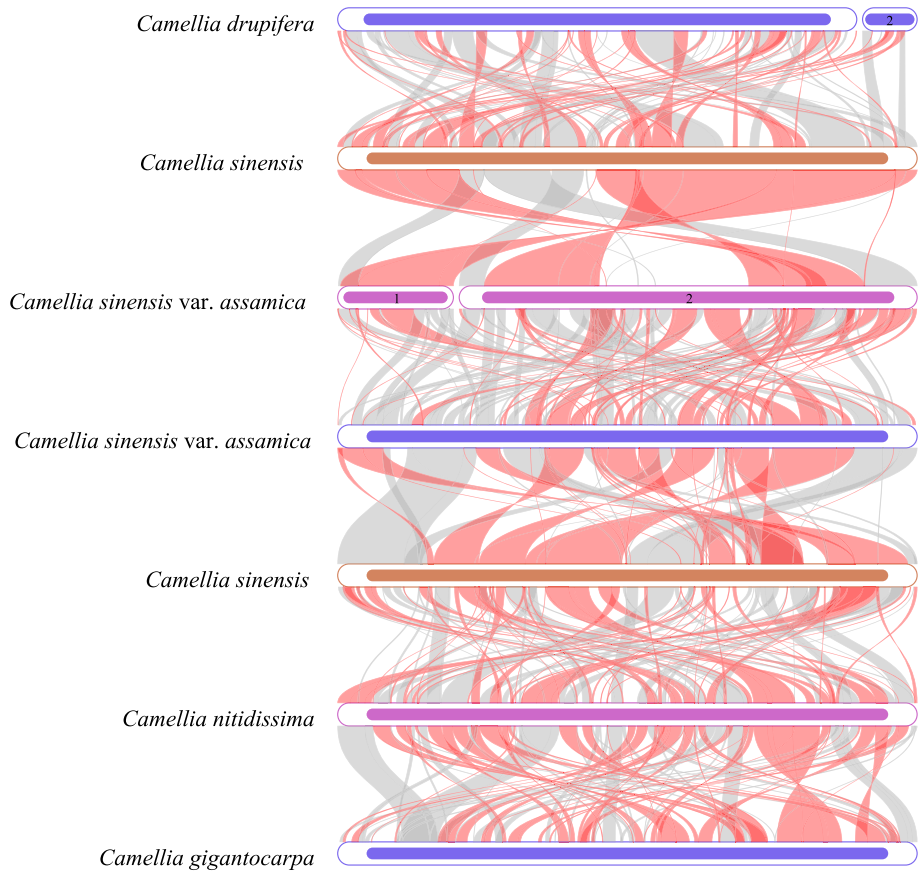

Fig. 9 Collinear analysis of seven *Camellia* species. The red arcs indicate inverted regions, while the gray arcs indicate better homologous regions

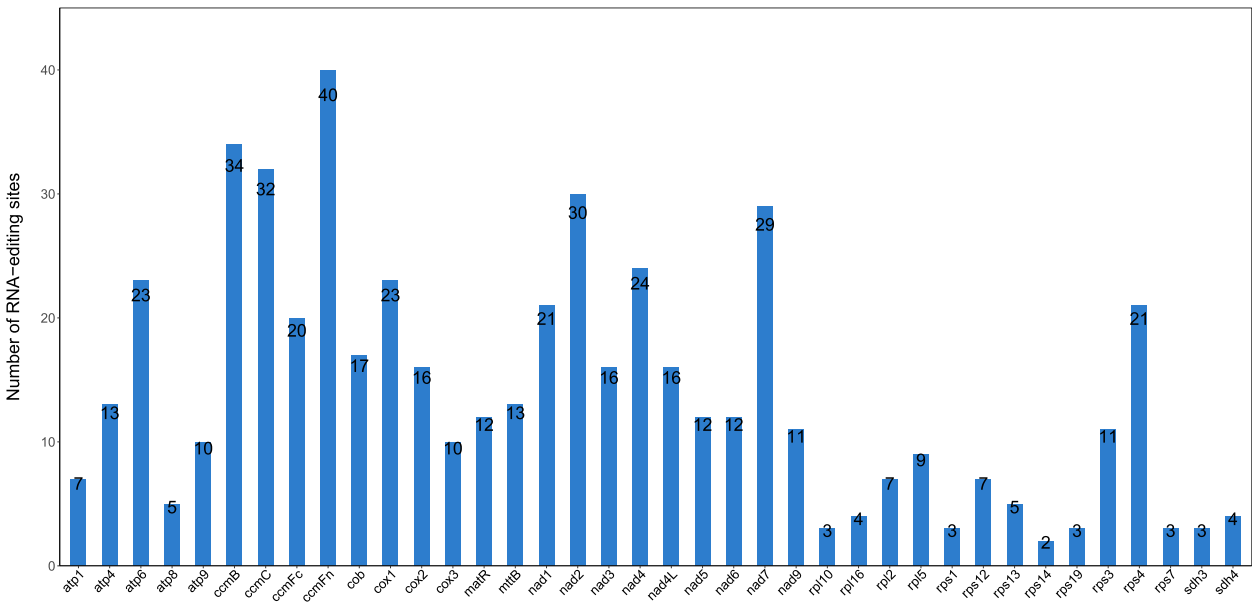

Fig. 10 Prediction of RNA editing sites based on the PCGs

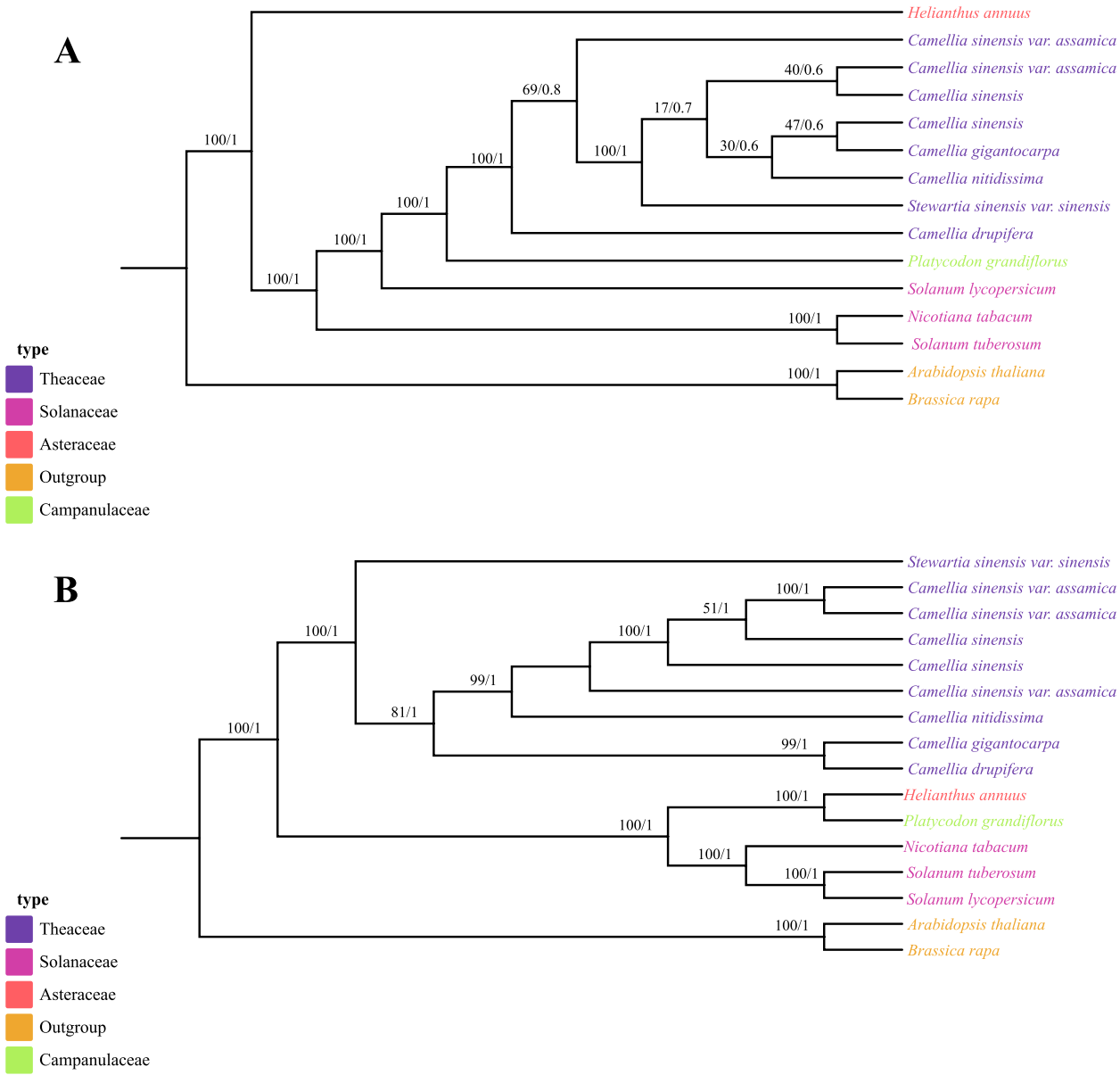

Fig. 11 Molecular phylogenetic analysis was conducted on 14 plant species using sequences from both mitochondrial and chloroplast genomes. **A** A phylogenetic tree was generated using conserved protein sequences and analyzed with Maximum Likelihood (ML) and Bayesian Inference (BI) methods. The reliability of the tree was evaluated with bootstrap scores from 1000 replicates, with ML bootstrap support values and BI posterior probabilities indicated at the corresponding nodes. **B** The tree was constructed using conserved protein sequences from the chloroplast genomes of the 14 plant species, applying the same methods as those used for the mitochondrial genome-based tree

Discussion

The first complete mitochondrial genome of Tea-oil Camellia species

Previous research appeared to resolve doubts regarding the structure of plant mitogenomes, suggesting that they are generally represented as a single circular molecule, without the presence of isoforms [43]. In the Tea-oil Camellia species, the assembly of organelle genomes are challenging due to complex structural rearrangements and high levels of repetitive sequences [26] Currently, only one partial mitochondrial genome of Tea-oil Camellia specie were reported, which was assembled into a circular model [44]. With the advent of advanced sequencing technologies, especially third-generation sequencing, the complexity of mitogenomes has become more apparent as more genomes are successfully

assembled [32]. Based on our study, we first confirmed a multi-branch mitochondrial conformation for Tea-oil Camellia species (*C. drupifera*). Although this structure has also been observed in other plant mitogenomes, such as *Picea sitchensis* [33] and *Coffea arabica* [45]. Due to the lack of comprehensive mitochondrial genome resources for Tea-oil Camellia species, a complete characterization of their mitochondrial genomes is currently difficult [24]. For example, we found that, with equivalent sequencing data, the mitochondrial content in *C. drupifera* was notably lower compared to other related species. Representing the *C. drupifera* mitochondrial genome as two linear molecules could facilitate comparisons with other Camellia species. Therefore, our findings not only provide valuable insights for the assembly of Tea-oil Camellia mitochondrial genomes but also lay the groundwork for future research into *Camellia* species and other related plants, providing an essential foundation for comparative genomics and evolutionary studies.

Mitochondrial Genome Features and Evolution of *C. drupifera*

So far, the mitochondrial genomes of *Camellia* species exhibit notable size differences, ranging from 707,441 bp to 1,081,996 bp. Among them, *C. sinensis* var. *assamica* is recorded as the largest, measuring 1,081,996 bp. Compared to another Tea-oil Camellia species, *C. gigantocarpa*, the length of *C. drupifera* mitochondrial genome is longer. This variation may be attributed to the fact that *C. drupifera* (2n = 7x, 8x, 10x, and 12x) is polyploid with fully sequenced mitochondrial genomes, whereas *C. gigantocarpa* (2n = 2x) is diploid, with only partially assembled mitochondrial genomes. Whole genome duplication (WGD) may lead to the creation of duplicate genes and the movement of genetic material in plant mitochondria, which may contribute to the expansion of the mitochondria genome [46–48].

The *C. drupifera* mitochondrial genome has 40 PCGs, 32 tRNAs, and 3 rRNAs. In the seven *Camellia* species, we found several instances of gene duplication, such as *atp9, rpl2, rps2,* and *rps11*. This redundancy could be a consequence of gene loss events in the evolutionary process of *Camellia*. The event of loss of genes in plant mitochondrial genomes will provide new novel perspective on genomic evolution that has yet to be explored uncovered [49].

Repeated sequences are an important feature of genomes, influencing genome evolution, inheritance, and variation [50]. It also plays an indispensable role in gene expression, transcriptional regulation, chromosome construction and physiological metabolism [51]. Beyond that, repeated sequences are crucial in promoting gene recombination within seed plant mitochondrial DNA, contributing to the expansion of the mitochondrial genome [52, 53]. The mitochondrial DNA of *C. drupifera* contains 269 SSRs, 43 tandem repeats, and 802 dispersed repeats, contributing to its complex and branched genomic structure. In our analysis of repeat regions, we found only a weak correlation between SSR frequency and mitochondrial genome size, which may be attributed to differences in evolutionary rates among *Camellia* species. While the mitochondrial genomes of *Camellia* species differ in size, their GC content and gene structure are relatively uniform, indicating the conserved in mitochondrial structure and function.

Research indicates that homologous fragments can dynamically transfer between chloroplast and mitochondrial genomes, emphasizing the interconnected and evolving nature of these genetic systems [54]. In *C. drupifera*, homologous fragments were detected across both chloroplast and mitochondrial genomes, comprising 1.73% of the total mitochondrial DNA. The repeated segments will provide new insight in *Camellia* evolution. Our examination of these fragments showed that seven tRNA genes, initially present in the chloroplast genome may have lost their original functionality or undergone changes to become pseudogenes. This result further demonstrated gene transfer often occurs between mitochondrial and chloroplast genomes in higher plants, which will caused the pseudogene loss or alteration in related genes [42].

Genome collinearity analysis is a method for comparing the similarity and co-evolution of genomic sequences across different species or within the same species, helping us understand the structure and evolutionary processes of genomes [55]. Collinearity analysis identified 45 conserved genes within aligned genomic regions, which play a significant role in the genetic diversity, evolutionary processes, and gene expression regulation in *Camellia* species [56, 57]. Furthermore, the arrangement of collinear blocks across mitochondrial genomes exhibited inconsistency, with multiple gene rearrangements observed in seven *Camellia* species, contributing to the reduction in the length of collinear blocks. This observation supports the idea that while the mitochondrial genomic arrangement was highly conserved among these seven *Camellia* species, they have also experienced frequent gene recombination events.

Phylogenetic analysis

Mitochondrial and chloroplast genomes offer high-resolution genetic data with conserved features and rapid evolutionary rates, which were used for reveal phylogenetic analysis [58, 59]. We conducted phylogenetic analysis of eight Theaceae species and seven other plant species using PCGs derived from both mitochondrial and

chloroplast genomes. The overall tree structure based on mitochondrial sequences was inconsistent with that constructed using shared genes from the chloroplast genomes. Based on the support rates of ML and BI tree, PCGs of chloroplast genome exhibited higher reliability. However, reconstructing the phylogeny of Theaceae remains challenging. For example, *C. drupifera* clustered with *C. gigantocarpa* based on chloroplast sequences, while *C. drupifera* formed a basal clade within Theaceae when using mitochondrial sequences. The reason is the variation in mitochondrial PCGs is smaller than that in chloroplast PCGs. In some *Camellia* species, certain mitochondrial PCGs exhibit no variation, which may lead to unreliable phylogenetic relationships.

The tribe *Stewartieae*, which is regarded as the earliest-diverging lineage within the Theaceae family, plays a crucial role in phylogenetic research aimed at gaining insights into the evolutionary development of the tea plant family [19]. In our study, the tree reconstructed using chloroplast sequences support this idea, whereas not in mitochondrial sequences. Moreover, the phylogenetic trees still could not fully resolve the phylogeny of *Camellia*. Due to the lack of genomic data from representative species, our ability to assess intergeneric relationships within Theaceae is limited. Therefore, obtaining and analyzing more mitochondrial and nuclear genomes is expected to yield a more complete understanding of the phylogenetic relationships among Theaceae species in the future. Currently, fewer than ten mitochondrial genomes of *Camellia* have been sequenced, which may introduce data bias and limit phylogenetic resolution.

Future direction

Nucleocytoplasmic interaction is the co-evolution process between nuclear genome and organelle genome [60, 61]. The process of nucleocytoplasmic interaction is complex and long-lasting, and plays an important role in cellular respiration, photosynthesis, lipid metabolism, and species differentiation [62–64]. Although numerous studies have suggested nucleocytoplasmic interactions, few have identified the specific nuclear genes and mitochondrial genetic variations involved within a single species [65]. However, there are no reports of the nuclear genome of *C. drupifera.* It is difficult to enhance the nucleocytoplasmic interaction between nuclear and mitochondrial within *C. drupifera*, such as cytoplasmic male sterility and evolutionary trajectories of organellar targeted genes.

The phylogeny of *Camellia* still remains controversial [66–68]. While reliable reference genome of *C. drupifera* will hopefully increase the reliability of speciation and evolution pattern of Tea-oil Camellia. In order to better understand the evolutionary history of Tea-oil Camellia and its closest relatives, the method of pan-genome inclusion of more taxa of *Camellia* will be crucial [34, 69–71].

Conclusions

The first complete mitogenome of main Tea-oil Camellia cultivar *C. drupifera* was successfully assembled, which exhibited a multi-branch structure composed of two linear molecules. A total of 24 core genes were found. The GC content of the mitochondrial DNA in *C. drupifera* was comparable to that of other *Camellia* species. The Ka/Ks analysis revealed that the *atp4* and *matR* genes are under positive selection. Additionally, the presence of gene transfer between organelles and conserved collinear blocks points to genome rearrangement and recombination, offering important insights into its genetic structure. RNA editing events may play a role in enhancing the stability of protein structures. The phylogenetic tree showed inconsistency and difficulties in inferring the phylogeny of Theaceae. This study will support the further exploration of population genetics and phylogeny in *Camellia* and other Theaceae members. Furthermore, we intend to include more samples from *Camellia* species and perform pan-genome analyses in future research.

Materials and methods

Plant material collection, DNA extraction, and sequencing

The young and healthy leaves of *C. drupifera* were obtained by a cutting seedling from the nursery of Hainan Academy of Agricultural Science (Haikou, Hainan, China). Liquid nitrogen was used to freeze the leaves. The total genomic DNA was extracted by Plant Genomic DNA Kit (Tiangen Biotech Co., Ltd., Beijing, China) following the manual. The extracted total DNA was evaluated by NanoDrop 2000 spectrophotometer (Thermo Scientific, USA) and stored at -20 °C until use. The complete mitochondrial genomes and chloroplast genomes of *C. drupifera* were obtained by the "3 + 2" strategy which sequenced by the long-reads obtained from the Nanopore sequencing platform and corrected by the short-reads using the Illumina Novaseq 6000 platform. A summary of the sequencing results of long-reads and short-reads were showed in Table S7 and Table S8, respectively.

Genome assembly and annotation

The assembly strategy is as follows: 1). We utilized Minimap2 (v.2.24) to align the Nanopore reads to our draft assembly of *C. drupifera* [72]. 2). The aligned reads were extracted and subjected to de novo assembly. 3). Initially, Flye (v.2.9.5) [73] was used to assemble the aligned data, followed by Racon v1.4.3 [74]. 4). After that, using Bowtie2 v2.5.4 [75] to align the short-reads to the previous

correction results, using Unicycler v0.5.1 [76] for mixed assembly, and 5). Split the GFA file according to the coverage of the long-reads to obtain the final assembly result.

The mitochondrial genomes were annotated by BlastN [77]. Mitochondrial genes were identified and queried against the NCBI database. Additionally, tRNA genes were detected using tRNA scan-SE software (v. 2.0.12) (http://lowelab.ucsc.edu/tRNAscan-SE/, accessed on 10 October 2023). The boundaries of the introns were manually reviewed and corrected to ensure the complete structure of the protein-coding genes. The newly sequenced mitochondrial genomes were deposited in GenBank under the accession numbers PQ041261 and PQ041262. Mitochondrial genome maps were constructed using the OGDRAW [78].

Comparative mitochondrial genomic analyses

The mitochondrial genomes of *C. sinensis* var. *assamica* cultivar Duntsa (OL989850), *C. sinensis* (MH376284.1), *C. sinensis* (OM809792.1), *C. sinensis* var. *assamica* (MK574876.1 and MK574877.1), *C. nitidissima* (ON645224), and *C. gigantocarpa* (OP270590, partial genome) were used to visualization and collinearity analysis by Mauve v2.4.1 software. The horizontal axis in each box represents the assembled sequences, while the vertical axis represents other sequences. The red lines within the boxes indicate forward alignments, and the blue lines represent reverse complement alignments. Non-Synonymous substitution rate, synonymous substitution rate and the ratio of Ka/Ks were calculated by KaKsCalculator2, *C. drupifera* mitochondrial genome as reference [79], and R package (ggplot2) plotted boxplots of paired Ka/Ks values). The protein-coding genes (PCGs) of *Camellia* species were also extracted by Phylosuite software (v1.2.2) [80]. Nucleotide diversity (Pi) values for shared PCGs were calculated by DnaSP v6.12.03 with a sliding window of 100 bp and a step size of 20 bp [81].

Analysis of repeat structures and SSRs

The tandem repeats of mitochondrial genome in C. *drupifera* were analyzed by the Tandem Repeats Finder v4.09 software (https://tandem.bu.edu/trf/trf.advanced.submit.html) with the parameters: 2, 7, 7, 80, 10, 50, 2000, -f, -d and -m [82]. The SSRs were identified by MISA (https://webblast.ipk-gatersleben.de/misa/) with the parameters: 10, 5, 4, 3, 3 and 3 [83]. Dispersed using blastn (v2.10.1 parameters: word_size 7, evalue e 1–5, remove redundant, removal of tandem repeat) software to identify. Using circos v0.69–5 to visualize them.

Codon Usage bias

MEGA software (v7.0) was employed to assess codon usage and determine RSCU values for the mitochondrial genome's protein-coding genes [84]. GC content of the coding genes was determined using the CUSP tool (https://www.bioinformatics.nl/cgi-bin/emboss/cusp). ENC values were calculated using CodonW to assess codon usage efficiency [85], and the ENC value represented the degree of random selection of genomic codon usage deviation.

Genomic synteny analysis

Using GetOrganelle, the chloroplast genome of *C. drupifera* was assembled and annotated with CPGAVAS2 [86]. The comparison of homologous sequences between chloroplast and mitochondrial genomes was performed using BLASTN with default parameters [87].

RNA editing prediction

The RNA editing sites within the shared protein-coding genes (PCGs) of *C. drupifera* were forecasted using PREP-M with a threshold score of C = 0.2 [88].

Phylogenetic analyses

Phylogenetic analyses involved a total of 15 species by shared PCGs from mitochondrial and chloroplast genomes, including two outgroup species. The optimal evolutionary model for the PCGs was determined using ModelTest-NG [89] based on AIC criteria. Maximum likelihood (ML) analyses were conducted for both datasets using RAxML-NG [90] with 1000 rapid bootstrap replicates. Phylogenetic trees were inferred using MrBayes v3.2.7 with the MCMC method over 1,000,000 generations, with sampling intervals of 100 generations and a burn-in of 25% of the total generations [91]. The FigTree v.1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/) program was utilized for the visualization of phylograms.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12870-024-05996-4.

Supplementary Material 1: Table S1 Summary of sequence data obtained from the Nanopore platform. Table S2 Summary of sequence data obtained from the Illumina platform. Table S3 The pi values in PCGs of *Cameilia*. Table S4 The statistics in SSRs of *Cameilia*. Table S5 The statistics in Tandem repeats of *C. drupifera*. Table S6 The statistics in Dispersed repeats of *C. drupifera*. Table S7 The GC contents and ENC of *C. drupifera*. Table S8 The statistics in Predicted RNA editing site of *C. drupifera*.

Acknowledgements

We thank Lei Xu (Genepioneer Biotechnologies Co. Ltd, Nanjing) for assistance with the data analysis.

Authors' contributions

D.Z. designed and supervised the project. H.L. wrote the manuscript. H.L. annotated and analyzed the genomes. H.Q., J.C., M.Y., Y.W., and X.S. prepared the samples and performed the experiments. C.W., T.X., X.F., S.F., and C.C.

analyzed the data. D.Z. and H.L. revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding
This study was supported by the National Natural Science Foundation of China (32360414), Hainan Province Science and Technology Special Fund (FW20230002), the Scientific and technological innovation team of Hainan Academy of Agricultural Sciences (HAAS2023TDYD05), the Basic scientific research business expenses of HAAS (ITH2024ZD02, HAAS2022KJCX01, HAAS2023RCQD13).

Data availability
The newly sequenced mitochondrial genomes were deposited in GenBank under the accession numbers PQ041261 and PQ041262.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Institute of Tropical Horticulture Research, Hainan Academy of Agricultural Sciences, Haikou 571100, China. [2]School of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai 200240, China. [3]Sanya Institute, Hainan Academy of Agricultural Sciences, Sanya 572025, China. [4]Key Laboratory of Tropic Special Economic Plant Innovation and Utilization, Haikou 571100, China. [5]National Germplasm Resource Chengmai Observation and Experiment Station, Chengmai 571100, China. [6]College of Life Science, Sichuan Agricultural University, Ya'an 625014, Sichuan Province, China. [7]School of Life Sciences, Technical University of Munich, Freising 85354, Germany.

Received: 22 August 2024 Accepted: 18 December 2024
Published online: 03 January 2025

References

1. Yao X, Ren H. Oil-tea Camellia Genetic Resource in China. Beijing: Science Press; 2022.
2. Qin S, Rong J, Zhang W, Chen J. Cultivation history of Camellia oleifera and genetic resources in the Yangtze River Basin. Biodiversity Science. 2018;26(4):384–95.
3. Wang X, Zeng Q, del Mar CM, Wang L. Profiling and quantification of phenolic compounds in Camellia seed oils: Natural tea polyphenols in vegetable oil. Food Res Int. 2017;102:184–94.
4. Zhuang R. Oil-Tea Camellia in China. Beijing: China Forestry Publishing House; 2008.
5. Ming T, Bartholomew B. Camellia In Flora of China. Beijing & St. Louispp: Science Press & Missouri Botanical Garden Press; 2007.
6. Xia T, Sun X, Chen J, Ma G, Wang C, Qi H, Feng Y, Liang Z, Zheng D. Leaves Phenotypic Diversity Analysis of Tea-oil Camellia Germplasm in Hainan Island. Molecular Plant Breeding. 2022;20(16):5484–94.
7. Yang L, Zhang Z, Chen JC. Xuan, Ji QZ, Daojun: The quantitative characters and diversity of oiltea fruit in Hainan province. Non-wood Forest Research. 2018;36(3):69–76.
8. Li C, Hou L. Review on volatile flavor components of roasted oilseeds and their products. Grain & Oil Science and Technology. 2018;1(4):151–6.
9. Wang J, Tang X, Chu Q, Zhang M, Zhang Y, Xu B. Characterization of the volatile compounds in Camellia oleifera seed oil from different geographic origins. Molecules. 2022;27(1):308.
10. Xia T, Xiong Z, Sun X, Chen J, Wang C, Chen Y, Zheng D. Metabolomic profiles and health-promoting functions of Camellia drupifera mature-seeds were revealed relate to their geographical origins using comparative metabolomic analysis and network pharmacology approach. Food Chem. 2023;426: 136619.
11. Yao G, Tang X, Ye Z, Yan W, Yu J, Wu Y, Zhang J, Yang D. Protective effect of Camellia vietnamensis active peptide on alcohol-induced hepatocyte injury. Food Hydrocolloids. 2021;32(1):425–49.
12. Qi H, Sun X, Wang C, Chen X, Yan W, Chen J, Xia T, Ye H, Yu J, Dai J. Geographic isolation causes low genetic diversity and significant pedigree differentiation in populations of Camellia drupifera, a woody oil plant native to China. Ind Crops Prod. 2023;192: 116026.
13. Qi H, Sun X, Yan W, Ye H, Chen J, Yu J, Jun D, Wang C, Xia T, Chen X: Genetic relationships and low diversity among the tea-oil Camellia species in Sect. Oleifera, a bulk woody oil crop in China. Front Plant Sci. 2020;13:996731.
14. Xu Z, Yuan D, Tang Y, Wu L, Zhao Y. Camellia hainanica (Theaceae) a new species from Hainan, supported from morphological characters and phylogenetic analysis. Pak J Bot. 2020;52(3):1025–32.
15. Liang H, Chen J. Comparison and Phylogenetic Analyses of Nine Complete Chloroplast Genomes of Zingibereae. Forests. 2021;12(6):710.
16. Liang H, Zhang Y, Deng J, Gao G, Ding C, Zhang L, Yang R: The Complete Chloroplast Genome Sequences of 14 Curcuma Species: Insights Into Genome Evolution and Phylogenetic Relationships Within Zingiberales. Front Genet. 2020;11.
17. Fuentes-Pardo AP, Ruzzante DE. Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. Mol Ecol. 2017;26(20):5369–406.
18. Holušová K, Čmejlová J, Suran P, Čmejla R, Sedlák J, Zelený L, Bartoš J: High-resolution genome-wide association study of a large Czech collection of sweet cherry (Prunus avium L.) on fruit maturity and quality traits. Hortic Res. 2023;10(1):uhac233.
19. Liu D, Zhang Z, Hao Y, Li M, Yu H, Zhang X, Mi H, Cheng L, Zhao Y. Decoding the complete organelle genomic architecture of Stewartia gemmata: an early-diverging species in Theaceae. BMC Genomics. 2024;25(1):114.
20. Lin P, Wang K, Wang Y, Hu Z, Yan C, Huang H, Ma X, Cao Y, Long W, Liu W. The genome of oil-Camellia and population genomics analysis provide insights into seed oil domestication. Genome Biol. 2022;23:1–21.
21. Gong W, Xiao S, Wang L, Liao Z, Chang Y, Mo W, Hu G, Li W, Zhao G, Zhu H. Chromosome-level genome of Camellia lanceoleosa provides a valuable resource for understanding genome evolution and self-incompatibility. Plant J. 2022;110(3):881–98.
22. Shen T, Huang B, Xu M, Zhou P, Ni Z, Gong C, Wen Q, Cao F, Xu L: The reference genome of Camellia chekiangoleosa provides insights into Camellia evolution and tea oil biosynthesis. Hortic Res. 2022;9:uhab083.
23. Fen Z, Feng L, Lin P, Jia JG, Lizhi: Chromosome-scale genome assembly of Camellia crapnelliana provides insights into the fatty acid biosynthesis. bioRxiv. 2024:2024.2001. 2007.574508.
24. Zhang L, Shi Y, Gong W, Zhao G, Xiao S, Lin H, Li Y, Liao Z, Zhang S, Hu G: The tetraploid Camellia oleifera genome provides insights into evolution, agronomic traits, and genetic architecture of oil Camellia plants. Cell Rep. 2024;43(11):114902.
25. Liang H, Qi H, Wang Y, Sun X, Wang C, Xia T, Chen J, Ye H, Feng X, Xie S *et al*: Comparative chloroplast genome analysis of Camellia oleifera and C. meiocarpa: phylogenetic relationships, sequence variation and polymorphism markers. Trop Plants. 2024:1–12.
26. Zhu H, Wang F, Xu Z, Wang GH, Lisong, Cheng J, Jin S, Ge XJ, Shuangxia: The complex hexaploid oil-Camellia genome traces back its phylogenomic history and multi-omics analysis of Camellia oil biosynthesis. Plant Biotechnol J. 2024:1–17.
27. Archibald JM. Endosymbiosis and eukaryotic cell evolution. Curr Biol. 2015;25(19):R911–21.
28. Gray MW, Burger G, Lang BF. Mitochondrial evolution. Science. 1999;283(5407):1476–81.
29. Wideman JG, Monier A, Rodríguez-Martínez R, Leonard G, Cook E, Poirier C, Maguire F, Milner DS, Irwin NA, Moore K. Unexpected mitochondrial genome diversity revealed by targeted single-cell genomics of heterotrophic flagellated protists. Nat Microbiol. 2020;5(1):154–65.
30. Butenko A, Lukeš J, Speijer D, Wideman JG. Mitochondrial genomes revisited: why do different lineages retain different genes? BMC Biol. 2024;22(1):15.

31. Zhou Q, Ni Y, Li J, Huang L, Li H, Chen H, Liu C. Multiple configurations of the plastid and mitochondrial genomes of Caragana spinosa. Planta. 2023;258(5):98.
32. Yang H, Ni Y, Zhang X, Li J, Chen H, Liu C. The mitochondrial genomes of Panax notoginseng reveal recombination mediated by repeats associated with DNA replication. Int J Biol Macromol. 2023;252: 126359.
33. Jackman SD, Coombe L, Warren RL, Kirk H, Trinh E, MacLeod T, Pleasance S, Pandoh P, Zhao Y, Coope RJ. Complete mitochondrial genome of a gymnosperm, Sitka spruce (Picea sitchensis), indicates a complex physical structure. Genome Biol Evol. 2020;12(7):1174–9.
34. Wang J, Kan S, Liao X, Zhou J, Tembrock LR, Daniell H, Jin S, Wu Z: Plant organellar genomes: Much done, much more to do. Trends Plant Sci. 2024.
35. Maréchal A, Brisson N. Recombination and the maintenance of plant organelle genome stability. New Phytol. 2010;186(2):299–317.
36. Palmer JD, Herbon LA. Plant mitochondrial DNA evolved rapidly in structure, but slowly in sequence. J Mol Evol. 1988;28:87–97.
37. Gianfranceschi L, Seglias N, Tarchini R, Komjanc M, Gessler C. Simple sequence repeats for the genetic analysis of apple. Theor Appl Genet. 1998;96:1069–76.
38. Gelfand Y, Rodriguez A, Benson G: TRDB—the tandem repeats database. Nucleic acids research. 2007;35(suppl_1):D80-D87.
39. Milligan BG, Hampton JN, Palmer JD. Dispersed repeats and structural reorganization in subclover chloroplast DNA. Mol Biol Evol. 1989;6(4):355–68.
40. Martin W, Herrmann RG. Gene transfer from organelles to the nucleus: how much, what happens, and why? Plant Physiol. 1998;118(1):9–17.
41. Rodríguez-Moreno L, González VM, Benjak A, Martí MC, Puigdomènech P, Aranda MA, Garcia-Mas J. Determination of the melon chloroplast and mitochondrial genome sequences reveals that the largest reported mitochondrial genome in plants contains a significant amount of DNA having a nuclear origin. BMC Genom. 2011;12:1–14.
42. Wang L, Liu X, Xu Y, Zhang Z, Wei Y, Hu Y, Zheng C, Qu X. Assembly and comparative analysis of the first complete mitochondrial genome of a traditional Chinese medicine Angelica biserrata (Shan et Yuan) Yuan et Shan. Int J Biol Macromol. 2024;257:128571.
43. Zhou P, Zhang Q, Li F, Huang J, Zhang M. Assembly and comparative analysis of the complete mitochondrial genome of Ilex metabaptista (Aquifoliaceae), a Chinese endemic species with a narrow distribution. BMC Plant Biol. 2023;23(1):393.
44. Lu C, Gao L-Z, Zhang Q-J. A high-quality genome assembly of the mitochondrial genome of the oil-tea tree Camellia gigantocarpa (Theaceae). Diversity. 2022;14(10):850.
45. Ni Y, Zhang X, Li J, Lu Q, Chen H, Ma B, Liu C: Genetic diversity of Coffea arabica L. mitochondrial genomes caused by repeat-mediated recombination and RNA editing. Front Plant Sci. 2023;14:1261012.
46. Fernandes Gyorfy M, Miller ER, Conover JL, Grover CE, Wendel JF, Sloan DB, Sharbrough J. Nuclear–cytoplasmic balance: whole genome duplications induce elevated organellar genome copy number. Plant J. 2021;108(1):219–30.
47. Ma X, Vanneste S, Chang J, Ambrosino L, Barry K, Bayer T, Bobrov AA, Boston L, Campbell JE, Chen H. Seagrass genomes reveal ancient polyploidy and adaptations to the marine environment. Nat Plants. 2024;10(2):240–55.
48. Liu S, Wu Z, Yang T, Xu J, Aishan S, Qin E, Ma K, Liu J, Qin R, Wang J. The Chrysosplenium sinicum genome provides insights into adaptive evolution of shade plants. Communications Biology. 2024;7(1):1004.
49. Feng Y, Wang Y, Lu H, Li J, Akhter D, Liu F, Zhao T, Shen X, Li X, Whelan J. Assembly and phylogenomic analysis of cotton mitochondrial genomes provide insights into the history of cotton evolution. The Crop Journal. 2023;11(6):1782–92.
50. Jurka J, Kapitonov VV, Kohany O, Jurka MV. Repetitive sequences in complex genomes: structure and evolution. Annu Rev Genomics Hum Genet. 2007;8:241–59.
51. Shapiro JA, von Sternberg R. Why repetitive DNA is essential to genome function. Biol Rev. 2005;80(2):227–50.
52. Gualberto JM, Newton KJ. Plant mitochondrial genomes: dynamics and mechanisms of mutation. Annu Rev Plant Biol. 2017;68(1):225–52.
53. Christensen AC. Plant mitochondrial genome evolution can be explained by DNA repair mechanisms. Genome Biol Evol. 2013;5(6):1079–86.
54. Waltz F, Salinas-Giegé T, Englmeier R, Meichel H, Soufari H, Kuhn L, Pfeffer S, Förster F, Engel BD, Giegé P. How to build a ribosome from RNA fragments in Chlamydomonas mitochondria. Nat Commun. 2021;12(1):7176.
55. Marsit S, Leducq J-B, Durand É, Marchant A, Filteau M, Landry CR. Evolutionary biology through the lens of budding yeast comparative genomics. Nat Rev Genet. 2017;18(10):581–98.
56. Hill MS, Vande Zande P, Wittkopp PJ. Molecular and evolutionary processes generating variation in gene expression. Nat Rev Genet. 2021;22(4):203–15.
57. Romero IG, Ruvinsky I, Gilad Y. Comparative studies of gene expression and the evolution of gene regulation. Nat Rev Genet. 2012;13(7):505–16.
58. Li-Zhen L, Dong-Yan T, Wu-Fu D, Shu-Dong Z. The chloroplast genome of Cephalanthera nanchuanica (Orchidaceae): comparative and phylogenetic analysis with other Neottieae species. BMC Genomics. 2024;25(1):1090.
59. Gould SB, Magiera J, García García C, Raval PK. Reliability of plastid and mitochondrial localisation prediction declines rapidly with the evolutionary distance to the training set increasing. PLoS Comput Biol. 2024;20(11):e1012575.
60. Dickinson H: Nucleocytoplasmic interaction. Ann Bot. 1987:61–73.
61. Sloan DB, Warren JM, Williams AM, Wu Z, Abdel-Ghany SE, Chicco AJ, Havird JC. Cytonuclear integration and co-evolution. Nat Rev Genet. 2018;19(10):635–48.
62. Millar AH, Heazlewood JL, Kristensen BK, Braun H-P, Møller IM. The plant mitochondrial proteome. Trends Plant Sci. 2005;10(1):36–43.
63. van Wijk KJ, Baginsky S. Plastid proteomics in higher plants: current state and future goals. Plant Physiol. 2011;155(4):1578–88.
64. Forsythe ES, Sharbrough J, Havird JC, Warren JM, Sloan DB. CyMIRA: the cytonuclear molecular interactions reference for Arabidopsis. Genome Biol Evol. 2019;11(8):2194–202.
65. Lian Q, Li S, Kan S, Liao X, Huang S, Sloan DB, Wu Z: Association analysis provides insights into plant mitonuclear interactions. Mol Biol Evol. 2024:msae028.
66. Zhao DW, Hodkinson TR, Parnell JA. Phylogenetics of global Camellia (Theaceae) based on three nuclear regions and its implications for systematics and evolutionary history. J Syst Evol. 2023;61(2):356–68.
67. Wu Q, Tong W, Zhao H, Ge R, Li R, Huang J, Li F, Wang Y, Mallano AI, Deng W. Comparative transcriptomic analysis unveils the deep phylogeny and secondary metabolite evolution of 116 Camellia plants. Plant J. 2022;111(2):406–21.
68. Zan T, He Y-T, Zhang M, Yonezawa T, Ma H, Zhao Q-M, Kuo W-Y, Zhang W-J, Huang C-H. Phylogenomic analyses of Camellia support reticulate evolution among major clades. Mol Phylogenet Evol. 2023;182: 107744.
69. Lian Q, Huettel B, Walkemeier B, Mayjonade B, Lopez-Roques C, Gil L, Roux F, Schneeberger K, Mercier R: A pan-genome of 69 Arabidopsis thaliana accessions reveals a conserved genome structure throughout the global species range. Nat Genet. 2024:1–10.
70. Liang Q, Muñoz-Amatriaín M, Shu S, Lo S, Wu X, Carlson JW, Davidson P, Goodstein DM, Phillips J, Janis NM: A view of the pan-genome of domesticated Cowpea (Vigna unguiculata [L.] Walp.). Plant Genome. 2024;17(1):e20319.
71. Wu D, Xie L, Sun Y, Huang Y, Jia L, Dong C, Shen E, Ye C-Y, Qian Q, Fan L. A syntelog-based pan-genome provides insights into rice domestication and de-domestication. Genome Biol. 2023;24(1):179.
72. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
73. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37(5):540–6.
74. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017;27(5):737–46.
75. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.
76. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol. 2017;13(6): e1005595.
77. Alverson AJ, Wei X, Rice DW, Stern DB, Barry K, Palmer JD. Insights into the evolution of mitochondrial genome size from complete sequences of Citrullus lanatus and Cucurbita pepo (Cucurbitaceae). Mol Biol Evol. 2010;27(6):1436–48.

78. Greiner S, Lehwark P, Bock R: OrganellarGenomeDRAW (OGDRAW) version 1.3. 1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic acids research* 2019;47(W1):W59-W64.
79. Wang D, Zhang Y, Zhang Z, Zhu J, Yu J: KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics, Proteomics and Bioinformatics* 2010;8(1):77–80.
80. Xiang C, Gao F, Jakovlić I, Lei H, Hu Y, Zhang H, Zou H, Wang G, Zhang D. Using PhyloSuite for molecular phylogeny and tree-based analyses. Imeta. 2023;2(1): e87.
81. Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sánchez-Gracia A. DnaSP 6: DNA sequence polymorphism analysis of large data sets. Mol Biol Evol. 2017;34(12):3299–302.
82. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27(2):573–80.
83. Beier S, Thiel T, Münch T, Scholz U, Mascher M. MISA-web: a web server for microsatellite prediction. Bioinformatics. 2017;33(16):2583–5.
84. Kumar S, Stecher G, Tamura K: MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular biology and evolution* 2016;33(7):1870–1874.
85. Sharp PM, Li W-H. Codon usage in regulatory genes in Escherichia coli does not reflect selection for 'rare'codons. Nucleic Acids Res. 1986;14(19):7737–49.
86. Shi L, Chen H, Jiang M, Wang L, Wu X, Huang L, Liu C. CPGAVAS2, an integrated plastome sequence annotator and analyzer. Nucleic Acids Res. 2019;47(W1):W65–73.
87. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19(9):1639–45.
88. Mower JP. PREP-Mt: predictive RNA editor for plant mitochondrial genes. BMC Bioinformatics. 2005;6:1–15.
89. Darriba D, Posada D, Kozlov AM, Stamatakis A, Morel B, Flouri T. ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. Mol Biol Evol. 2020;37(1):291–4.
90. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 2019;35(21):4453–5.
91. Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP: MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology* 2012;61(3):539–542.

Publisher's Note