Postprint: Characterization of the Chloroplast Genome of the Rare and Endangered Vietnamese Small-flowered Golden Camellia

Authors: Deng Yongbiao, Zhang Jin, Lan Lunli, Zhao Bo

Date: 2023-12-26T00:00:00+00:00

Abstract

Camellia minima is a rare and endangered species within the golden camellia group (sect. Chrysantha), for which no chloroplast genome studies have been reported to date. This study employed the Illumina HiSeq 2000 platform to perform sequencing, assembly, annotation, and analysis of the chloroplast genome of C. minima. The results demonstrated that: (1) The chloroplast genome of C. minima has a total length of 156,961 bp, exhibits a typical quadripartite structure, and contains 136 annotated genes, including 87 protein-coding genes, 41 tRNA genes, and 8 rRNA genes. (2) A total of 66 SSR loci and 39 repeat sequences were identified. (3) Codon usage bias favored codons ending with A/U. Integrated ENC-plot, PR2-plot, and neutrality analysis indicated that natural selection is the predominant factor influencing codon usage patterns. (4) Boundary analysis revealed variations in the length and position of the ycf1 gene among different species within the golden camellia group. (5) Phylogenetic analysis based on published chloroplast genomes of golden camellia group plants showed that C. minima and C. micrantha form a monophyletic clade, indicating a close phylogenetic relationship. These findings provide important reference information for exploring species evolution and improving exogenous gene expression, while also establishing a theoretical foundation for the conservation and utilization of golden camellia group plants.

Full Text

Preamble

DOI: 10.11931/guihaia.gxzw202304069

Analysis of Chloroplast Genome Characteristics of the Rare and Endangered Plant *Camellia minima*

DENG Yongbiao, ZHANG Jin, LAN Lunli, ZHAO Bo*

(Department of Pharmacognosy, School of Pharmacy, Guilin Medical University, Guilin 541199, Guangxi, China)

Abstract

Camellia minima is a rare and endangered species within Sect. *Chrysantha*, yet no studies on its chloroplast genome have been reported to date. This study sequenced, assembled, annotated, and analyzed the chloroplast genome of *C. minima* using the Illumina HiSeq 2000 platform. The results showed that: (1) The chloroplast genome of *C. minima* is 156,961 bp in length, exhibits a typical quadripartite structure, and contains 136 annotated genes, including 87 protein-coding genes, 41 tRNA genes, and 8 rRNA genes. (2) Sixty-six SSR loci and 39 repetitive sequences were identified. (3) Codon usage bias favored A/U-ending codons, and comprehensive analyses including ENC-plot, PR2-plot, and neutrality analysis suggested that natural selection is the dominant factor shaping codon usage patterns. (4) Boundary analysis revealed variations in the length and position of the *ycf1* gene among different yellow camellia species. (5) Phylogenetic analysis of published chloroplast genomes from Sect. *Chrysantha* indicated that *C. minima* clusters with *C. micrantha* with strong support, suggesting a close phylogenetic relationship. These findings provide important reference information for exploring species evolution and enhancing exogenous gene expression, while establishing a theoretical foundation for the conservation and utilization of Sect. *Chrysantha* plants.

Keywords: *Camellia minima*, chloroplast genome, characteristic analysis, phylogenetic analysis, codon preference
Chinese Library Classification: Q943
Document Code: A

Introduction

Sect. *Chrysantha* belongs to the family Theaceae and genus *Camellia* L., primarily distributed in Guangxi and Yunnan provinces of China, as well as parts of Vietnam (Sai, 2018). These plants typically inhabit karst landscapes and humid mountainous areas with minimal human disturbance (Li et al., 2022). All species within Sect. *Chrysantha* are listed as second-class nationally protected wild plants in China (National Forestry and Grassland Administration, 2021). Known as the "giant panda of the plant kingdom" and "queen of teas," this section represents the only group in *Camellia* with golden-yellow petals (Zhang et al., 2018; Wu et al., 2020). The petals and leaves are rich in flavonoids and are commonly used as health foods and beverages (Liu et al., 2021). Frequent interspecific hybridization and polyploidization have made taxonomic studies of these plants extremely challenging (Zhang et al., 2019). Wei et al. (2022) reconstructed the phylogeny of 20 Chinese yellow camellia species using ddRAD-seq, transcriptome, and nrITS data, combining morphological evidence to determine interspecific relationships. Their study revealed strong hy-

bridization/introgression signals, indicating that reticulate evolution is the primary cause of incongruence between nuclear gene data and chloroplast genome data. However, that study only utilized the relatively small SSC region of the chloroplast genome. Accumulation of more complete chloroplast genome data will provide additional molecular evidence for future phylogenetic and reticulate evolutionary studies of *Camellia*, while also supporting the conservation and comprehensive utilization of Sect. *Chrysantha* plants.

The quadripartite structure is a typical feature of chloroplast genomes, consisting of one large single-copy (LSC) region, one small single-copy (SSC) region, and two inverted repeat (IR) regions, encoding 110–130 genes with a total size ranging from 120 to 180 kb (Li et al., 2021). Compared with nuclear genomes, chloroplast genomes are more stable in gene structure, content, and sequence, with slower evolutionary rates, making them widely applicable in phylogenetic studies, DNA barcoding, genetic engineering, and population genetics (Dong et al., 2018). In recent years, the cost of chloroplast genome sequencing has decreased dramatically, enabling successful application of increasing amounts of chloroplast genome data in plant phylogeny and evolution studies. Additionally, chloroplast genomes contain abundant repetitive sequences that serve as important resources for studying evolutionary processes and genetic characteristics (Hui et al., 2014). Simple sequence repeats (SSRs), also known as microsatellites, can serve as effective molecular markers for detecting population polymorphism and are widely used in molecular-assisted breeding and species conservation (Cavalier, 2002).

Codons serve as the link between nucleic acids and proteins, playing a crucial role in genetic information transfer (Liu et al., 2012). Studying codon usage patterns and identifying optimal codons helps design gene expression vectors to improve target gene expression, holding significant applied value in variety improvement (Qi et al., 2015; Hu et al., 2019).

Camellia minima, a species within Sect. *Chrysantha*, is used as an ornamental plant and grafting rootstock (George & Anthony, 2015). Currently, no studies on the chloroplast genome of *C. minima* have been reported, limiting understanding of its genetic characteristics and phylogenetic relationships. Therefore, this study performed whole chloroplast genome sequencing and assembly of *C. minima* using high-throughput sequencing technology to address the following scientific questions: (1) Characterize the chloroplast genome map and sequence features of *C. minima*; (2) Analyze codon usage bias in the chloroplast genome and infer the dominant factors influencing codon usage patterns; (3) Investigate differences in base distribution at IR/SC boundary regions between *C. minima* and its close relatives; and (4) Clarify the phylogenetic position of *C. minima* within Sect. *Chrysantha*. This study provides an important theoretical foundation for subsequent research on species identification, genetic diversity analysis, chloroplast genetic engineering, and molecular breeding of *C. minima*, while also contributing richer chloroplast genome data for phylogenetic studies of Sect. *Chrysantha*.

Materials and Methods

1.1 Plant Material

Camellia minima was collected from the Camellia Garden in Nanning, Guangxi Zhuang Autonomous Region, China (108°21 43 E, 22°49 30 N). Fresh young leaves were sampled from a single individual, stored in self-sealing bags with silica gel at low temperature for subsequent total DNA extraction. Voucher specimens were deposited at the Guangxi Institute of Botany, Chinese Academy of Sciences (voucher No. CSF2020003).

1.2 DNA Extraction and Sequencing

Total DNA was extracted from dried leaves of *C. minima* using a modified CTAB method. Chloroplast genome sequencing services were provided by Beijing Genomics Biotechnology Co., Ltd. Raw sequencing data were processed using fastp software to remove sequences with excessive N bases, adapter sequences, and overly short reads, yielding 1,005.18 Mb of clean data for subsequent chloroplast genome assembly (Gu et al., 2018).

1.3 Genome Assembly, Annotation, and Sequence Characterization

Clean data were assembled using GetOrganelle v1.7.6.1 with the embplant_{pt} (land plant chloroplast) reference database, a maximum cycle extension of 20, and K-mer values of 21, 45, 65, 85, 105, and 127 called by Spedas. Assembly results were visualized using Bandage software (Jian et al., 2018). Annotation was performed using the online tool CPGAVAS2 (http://47.96.249.172:16019/analyzer/home) with *C. parvipetala* (NC_{067089}.1) as the reference genome. GB2sequin was used to check for inverted sequence partitions and manually adjust accurate positions to obtain complete annotation results (Shi et al., 2019; Pascal & Stephan, 2018). Finally, the online tool CPGview (http://www.1kmpg.cn/cpgview/) was used to draw the chloroplast genome circular map (Liu et al., 2023). The annotated chloroplast genome was submitted to GenBank under accession number NC_{069310}.1, with corresponding SRA, BioProject, and BioSample numbers SRR20648317, PRJNA861872, and SAMN29930849, respectively.

1.4 Codon Usage Analysis

1.4.1 Calculation of Codon-Related Parameters After removing duplicate genes, sequences containing stop codons, and coding sequences (CDS) shorter than 300 bp, 52 CDS sequences were selected for codon analysis. Relative synonymous codon usage (RSCU) was analyzed using CodonW 1.4.2 software. The EMBOSS toolkit was used to calculate total GC content and GC content at the first, second, and third codon positions (GC1, GC2, GC3) via the CUSP program, while the CHIPS program was used to analyze effective number of codons (ENC) (Zhu et al., 2022). Finally, the cor() function in R software was used to calculate correlations.

1.4.2 Neutrality Plot, ENC-Plot, and PR2-Plot Analysis In neutrality plot analysis, GC12 versus GC3 scatter plots were constructed to investigate correlations among nucleotides at the three codon positions and assess the effects of mutational pressure and natural selection on codon usage bias (Wei et al., 2014). GC12 represents the average of GC1 and GC2. Significant correlation between GC12 and GC3 (with R^2 approaching 1) indicates that mutational pressure is the primary determinant of codon usage bias. Conversely, non-significant correlation with a low or near-zero regression slope suggests that natural selection dominates codon preference (Sueoka, 1988).

ENC-plot analysis further examined the influence of base composition on codon usage bias by plotting GC3 on the x-axis and effective number of codons (ENC) on the y-axis, calculated as: ENC = 2 + GC3 + 29/[$GC3^2$ + (1-GC3)^2] (Yang et al., 2021; Xin et al., 2021). When codon usage bias is influenced solely by mutation, genes will distribute along or near the standard curve. Genes falling below the standard curve indicate that natural selection and other factors are the main influences (Chakraborty et al., 2020).

Parity rule 2 plot (PR2-plot) analysis was used to calculate the content of A, T, C, and G at the third codon position to estimate the effects of mutational pressure and natural selection (Xiang et al., 2015). PR2-plots were constructed with A3/(A3+T3) on the y-axis and G3/(G3+C3) on the x-axis. The central point (A=T, G=C) indicates no bias between natural selection and mutational pressure. Even distribution of genes around the central point suggests that codon preference may be entirely caused by mutational pressure; otherwise, codon usage may be influenced by natural selection and other factors (Xiang et al., 2015).

1.5 Repeat Sequence Analysis

Dispersed repeat sequences were analyzed using the online tool REPuter with the following parameters: maximum computed repeats = 200; minimum repeat length > 30 bp; sequence identity > 90%; Hamming distance = 3 (Kurtz et al., 2001). Simple sequence repeats were identified and statistically analyzed using the online tool MISA with minimum repeat thresholds of: 1–10, 2–5, 3–4, 4–3, 5–3, and 6–3 for each unit size, and a minimum distance of 100 bp between adjacent SSRs (Beier et al., 2017).

1.6 Chloroplast Genome IR Boundary Analysis

Based on phylogenetic analysis results, IRscope was locally deployed to visualize the LSC/IRb/SSC/IRa boundaries of *C. minima* and related species using *C. fascicularis* as the reference (Amiryousefi et al., 2018).

1.7 Phylogenetic Analysis

A phylogenetic tree was constructed using the complete chloroplast genome sequence of *C. minima* generated in this study and 23 other Sect. *Chrysantha*

chloroplast genomes obtained from NCBI, with *Polyspora penangensis* as the outgroup. The workflow included: multiple sequence alignment using MAFFT (version 7.505), removal of low-quality alignment regions using trimAl (V1.4) to improve quality and accuracy, and phylogenetic tree construction using IQ-TREE2 software with the maximum likelihood (ML) method under the TVM+F+I+I+R6 model (Katoh & Standley, 2003; Minh et al., 2021).

Results

2.1 Chloroplast Genome Characteristics

The chloroplast genome of *C. minima* is a circular double-stranded quadripartite structure (Figure 1 [Figure 1: see original paper]) with a total length of 156,961 bp. The genome contains two inverted repeat (IR) regions of 26,047 bp each, separated by a large single-copy (LSC) region of 86,600 bp and a small single-copy (SSC) region of 18,267 bp. The overall GC content is 37.32%, with 35.33% in the LSC region, 30.60% in the SSC region, and 42.98% in the IR region.

The chloroplast genome contains 136 functional genes (Table 1), including 87 protein-coding genes, 41 tRNA genes, and 8 rRNA genes. Among these, 78 genes are related to self-replication, 45 to photosynthesis, and 13 have unknown functions. The genome includes 16 duplicated genes (*ndhB*, *rpl2*, *rpl23*, *rps7*, *rps12*, *rrn4.5S*, *rrn5S*, *rrn16S*, *rrn23S*, *trnL-CAA*, *trnN-GUU*, *trnR-ACG*, *trnV-GAC*, *ycf2*, *ycf15*) and two quadruplicated genes (*trnI-GAU*, *trnA-UGC*). Ten genes contain introns: *petB*, *petD*, *rps16*, *rpl16*, *rpl12*, *atpF*, *rpoC1*, *clpP*, *ndhB*, and *ndhA*, with both *rpl12* and *ndhB* containing two introns.

2.2 Repeat Sequence Analysis

Analysis using the MISA online tool identified 66 SSR loci in the chloroplast genome of *C. minima*, including mononucleotide, dinucleotide, trinucleotide, tetranucleotide, and hexanucleotide repeats, with mononucleotide repeats being the most abundant (Figure 2 [Figure 2: see original paper]:C). These SSRs were distributed as follows: 17 in protein-coding regions, 40 in intergenic regions, and 9 in intron regions (Figure 2:B). REPuter analysis detected 39 dispersed repeat sequences, comprising 16 forward repeats and 23 palindromic repeats; no reverse or complementary repeats were found. These repeats ranged from 30 to 64 bp in length, with most located in the *ycf2* gene within the IR region (Figure 2:D).

2.3.1 Relative Synonymous Codon Usage Analysis

Analysis of 52 protein-coding sequences longer than 300 bp in the chloroplast genome revealed 30 codons with RSCU values >1, of which 13 ended with A and 16 ended with U, while only one ended with G. Among the 32 codons with RSCU values <1, most ended with C (16) or G (13), indicating a preference for A/U-ending codons (Figure 3 [Figure 3: see original paper]).

2.3.2 Neutrality Plot, ENC-Plot, and PR2-Plot Analysis

Neutrality plot analysis (GC12 vs. GC3) of chloroplast genes showed relatively concentrated but not dense distribution (Figure 4 [Figure 4: see original paper]:A). Mean GC12 and GC3 values were 43.10% and 27.49%, respectively, with a correlation coefficient of r=0.118 (R^2=0.014) and regression slope of 0.107, indicating no significant correlation between GC12 and GC3. This suggests that codon usage bias is minimally affected by mutational pressure and more strongly influenced by natural selection and other factors.

To assess codon bias in protein-coding genes, ENC values were calculated and analyzed. As shown in Figure 4B, most genes had ENC values below expected values, falling below the standard curve, indicating that codon usage bias is primarily influenced by natural selection rather than mutational pressure. ENC ratio frequency distribution ranged from -0.05 to 0.45 (Table 2), with nine genes (17.00%) distributed within -0.05 to 0.05 (near the standard curve) and the majority located farther away, confirming that natural selection has a greater influence on codon preference.

If codon usage were completely mutation-driven, nucleotides A, T, C, and G at the third codon position would be used equally. In this study, PR2-plot analysis showed non-uniform distribution across four quadrants (Figure 4:C), with most genes located in the lower half, particularly the lower right quadrant. Base usage frequency analysis revealed T>A and G>C, indicating imbalanced A/T and G/C codon usage preference, suggesting that codon usage bias in *C. minima* is influenced by both mutational pressure and natural selection.

2.4 IR Boundary Analysis

Chloroplast genomes have four boundaries between LSC, IRb, IRa, and SSC regions: LSC/IRb, IRb/SSC, SSC/IRa, and IRa/LSC. Comparative analysis of these boundaries in *C. minima* and six related yellow camellia species revealed only slight IR region expansion and contraction, with lengths ranging from 25,996 bp (*C. pubipetala*) to 26,096 bp (*C. perpetua*). At the LSC/IRb boundary, the *rps19* gene crossed into the IRb region by 46 bp in all species. At the IRb/SSC boundary, all seven species showed boundaries located at the overlap between the *ycf1* pseudogene and *ndhF* gene, with *ndhF* crossing the boundary by 39 bp in *C. chrysantha* and 25 bp in *C. pubipetala*. At the SSC/IRa boundary, the *ycf1* gene crossed this boundary, containing 1,034–1,055 bp within the IRa region. At the IRa/LSC boundary, all species showed boundaries located between the *rps19* copy and *trnH* gene, with *trnH* positioned 1 bp from the IRa/LSC boundary.

2.5 Phylogenetic Analysis

A phylogenetic tree was constructed based on complete chloroplast genome sequences from 25 Theaceae species (24 Sect. *Chrysantha* species and one *Polyspora* species), with *P. penangensis* as the outgroup using the maximum

likelihood method. The results showed that 24 Sect. *Chrysantha* species were divided into two main clades (Clade I and Clade II). Within Clade I, *C. minima* and *C. micrantha* formed a single branch with 100% bootstrap support, indicating the closest phylogenetic relationship between these two species (Figure 6 [Figure 6: see original paper]).

Discussion and Conclusion

All species within Sect. *Chrysantha* are second-class nationally protected plants in China (National Forestry and Grassland Administration, 2021) with high ornamental and medicinal value. *Camellia minima*, native to northern Vietnam, grows in humid, shaded valleys. Due to its strong adaptability, it is considered an excellent grafting rootstock and has attracted widespread attention in horticulture (Li et al., 2022). The chloroplast genome of *C. minima* exhibits a typical circular double-stranded quadripartite structure, 156,961 bp in length, with a GC content of 37.32%. The complete chloroplast genome annotation identified 136 genes, including 87 protein-coding genes, 41 tRNA genes, and 8 rRNA genes. Ten genes contain introns, with *rpl12* and *ndhB* each containing two introns, consistent with other Sect. *Chrysantha* plants and likely related to the special structure and replication mechanism of chloroplast genomes (Henry et al., 2016; Wu et al., 2021). Additionally, the chloroplast genome length, gene type and order, and GC content are similar to other published *Camellia* species (e.g., *C. japonica*, *C. oleifera*, *C. impressinervis*), suggesting conservative and slow evolution of the *C. minima* chloroplast genome (Ding et al., 2022; Sophiarani et al., 2019).

Simple sequence repeats (SSRs) arise from DNA strand slippage and are widely distributed in both nuclear and chloroplast genomes. SSRs typically have higher mutation rates than other neutral DNA regions. Due to their non-recombinant, haploid, and uniparentally inherited characteristics, SSRs serve as valuable genetic markers for plant population genetics, ecology, and evolutionary studies (Gui et al., 2020; Aii et al., 1997). This study identified 66 SSR loci in the chloroplast genome of *C. minima*, primarily located in the LSC region. All mononucleotide repeats consisted of A/T, similar to findings in other Sect. *Chrysantha* chloroplast genomes (Ding et al., 2022) and other angiosperms (Yang et al., 2014; Hui et al., 2014), likely due to higher structural stability of polyA and polyT compared with polyC and polyG (Jin et al., 2023). However, the number of SSR loci detected differed from Hui et al. (2014), possibly due to different parameter settings for SSR detection. The abundant cp-SSRs in *C. minima* can be used to detect population genetic polymorphism. Additionally, dispersed repeats in the chloroplast genome were mainly forward and palindromic repeats, similar to *C. tienii* (Ding et al., 2022). These repeats are important genetic resources that play significant roles in phylogenetic studies (Wei et al., 2022).

Codon usage bias is important for studying molecular evolution and heterologous gene expression (Li et al., 2021). This study provides the first systematic analysis of codon usage patterns in the *C. minima* chloroplast genome, revealing

that arginine (Arg), leucine (Leu), and serine (Ser) are the most abundant amino acids encoded by six codons each, while tryptophan (Trp) and methionine (Met) are encoded by only one codon each. The chloroplast genes preferentially use A/U-ending codons, consistent with codon analysis results for *C. nitidissima* (Geng et al., 2022). Neutrality plot analysis revealed no significant correlation between GC12 and GC3, with relatively concentrated gene distribution, indicating that natural selection is the main factor influencing codon preference (Zhang et al., 2012). Combined ENC-plot and PR2-plot analyses further demonstrated that codon usage bias in *C. minima* is influenced by multiple factors including mutational pressure, base composition, and gene length, with natural selection being the dominant factor, consistent with Ding et al. (2023) and Li et al. (2022). The finding that natural selection is the main driving force in chloroplast genome evolution deepens our understanding of the evolutionary history of *C. minima*, particularly regarding natural selection-related evolution. Codon preference analysis facilitates codon optimization and provides a theoretical basis for future transgenic technology in Sect. *Chrysantha* plants (Duan & Zhang, 2020).

In chloroplast genomes, IR contraction and expansion occur frequently, leading to pseudogene formation, gene duplication, and gene loss, which further cause positional changes at IR/SC junctions responsible for length variation in higher plant chloroplast genomes (Li et al., 2021). Comparative analysis of chloroplast genome boundaries in *C. minima* and six related Sect. *Chrysantha* species showed consistent IR region lengths (25,996–26,096 bp) across all seven species without gene loss, suggesting that high conservation of the IR region is crucial for maintaining its length and structural stability. Boundary shift analysis revealed differences in *ycf1* gene length and position among different Sect. *Chrysantha* species, identifying it as a potential mutational hotspot. Previous studies have recommended the *ycf1* gene as a core DNA barcode due to its high polymorphism; however, whether this highly variable region can serve as an effective DNA barcode for Sect. *Chrysantha* requires further validation (Li et al., 2021). Additionally, relatively conserved chloroplast genome boundaries among *C. minima* and its six relatives indicate close phylogenetic relationships, supported by subsequent phylogenetic analysis.

To date, numerous molecular markers—including RAPD, cpDNA *trnL-trnF*, AFLP, ISSR, and nrITS—have been used to elucidate relationships within *Camellia* (Ju et al., 2021). Although Wei et al. (2022) integrated ddRAD, transcriptome, nrITS, and SSC markers to study Sect. *Chrysantha* phylogeny, the chloroplast genome SSC region-based phylogenetic tree showed extremely low support values. This study analyzed the phylogenetic relationships of *C. minima* with 23 other Sect. *Chrysantha* species based on complete chloroplast genome sequences. The phylogenetic tree divided into two major clades, with *C. minima* and *C. micrantha* forming a highly supported branch, indicating the closest relationship. High support values across all branches demonstrate that complete chloroplast genome data provide important support for reconstructing phylogenetic relationships within Sect. *Chrysantha*.

In summary, this study presents the first complete chloroplast genome sequencing, assembly, annotation, and basic information analysis of *C. minima*, revealing its fundamental characteristics. By dissecting codon usage patterns and factors influencing codon preference, we identified high-frequency codons in the *C. minima* chloroplast genome, providing theoretical support for future transgenic research in Sect. *Chrysantha*. Furthermore, phylogenetic analysis based on complete chloroplast genomes clarified the systematic position of *C. minima* within Sect. *Chrysantha*. This study establishes a theoretical foundation for subsequent conservation and rational development and utilization research on *C. minima* and other Sect. *Chrysantha* plants.

References

AII J, KISHIMA Y, MIKAMI T et al., 1997. Expansion of the IR in the chloroplast genomes of buckwheat species is due to incorporation of an SSC sequence that could be mediated by an inversion[J]. Curr Genet, 31: 276-279.

AMIRYOUSEFI A, HYVÖNEN J, POCZAI P., 2018. IRscope: an online program to visualize the junction sites of chloroplast genomes[J]. Bioinformatics, 34(17): 3030-3031.

BEIER S, THIEL T, MüNCH T, et al., 2017. MISA-web: a web server for microsatellite prediction[J]. Bioinformatics, 33(16): 2583-2585.

CAVALIER ST, 2002. Chloroplast evolution: secondary symbiogenesis and multiple losses[J]. Curr Biol, 12: 62-64.

CHAKRABORTY S, YENGKHOM S, UDDIN A, 2020. Analysis of codon usage bias of chloroplast genes in *Oryza* species[J]. Planta, 252: 1-20.

DING XQ, BI YY, CHEN JT, et al., 2022. Analysis of the chloroplast genome characteristics of *Camellia tienii*[J]. Agric Sci Jiangsu, 50(23): 33-40.

DING XQ, CHEN SY, CHEN JT, et al., 2023. Codon bias analysis of 11 yellow *Camellia* chloroplast genome[J]. J Fujian Agric For Univ(Nat Sci Ed), 52(4): 1-9.

DING XQ, LI WF, WU JL, et al., 2022. Chloroplast genome characteristics and genetic relationship of yellow *Camellia*[J]. J Fujian Agric For Univ (Nat Sci Ed), 52(03): 1-11.

DONG W, XU C, WU P, et al., 2018. Resolving the systematic positions of enigmatic taxa: Manipulating the chloroplast genome data of Saxifragales[J]. Mol Phylogenet Evol, 126: 321-330.

DUAN YZ, ZHANG K, 2020. Comparative analysis and phylogenetic evolution of the complete chloroplast genome of *Ammopiptanthus*[J]. Acta Bot Boreal-Occident Sin, 40(8): 1323-1332.

GENG XS, JIA W, CHEN JN, et al., 2022. Codon usage bias analysis of

chloroplast genome in *Camellia nitidissima* [J]. Mol Plant Breed, 20(7): 2196-2203.

GEORGE O, ANTHONY SC, 2015. In pursuit of hidden camellias: 32 new camellia species from Vietnam and China [M]. 2nd edition. Sydney: Theaceae Exploration Associates.

GUI L, JIANG S, XIE D, et al., 2020. Analysis of complete chloroplast genomes of curcuma and the contribution to phylogeny and adaptive evolution[J]. Gene, 732:144355.

GU J, CHEN S, ZHOU Y, et al., 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor[J]. Bioinformatics, 34(17): i884-i890.

HENRY D, LIN CS, YU M, et al., 2016. Chloroplast genomes: diversity, evolution, and applications in genetic engineering[J]. Genome Biol, 17(134): 1-29.

HU XY, XU YQ, HAN YZ, et al., 2019. Codon usage bias analysis of the chloroplast genome of *Ziziphus jujuba* var. *spinosa*[J]. J For Environ, 39(6): 621-628.

HUI H, CHAO S, YUAN L, et al., 2014. Thirteen camellia chloroplast genome sequences determined by high-throughput sequencing: genome structures and phylogenetic relationships[J]. BMC Evol Biol, 14(1): 151-151.

JIAN JJ, YU WB, YANG JB, et al., 2018. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes[J]. Genome Biol, 21(1): 241.

JIN GZ, LI WJ, SONG F, et al., 2023. Comparative analysis of complete chloroplast genomes of *Artemisia* subgenus *Seriphidium* (Asteraceae: Anthemideae): insights into structural divergence and phylogenetic relationships[J]. BMC Plant Biol, 136: l-23.

JU NG, HOANG DKD, KIM CK, et al., 2021. Complete chloroplast genomes shed light on biogeography, divergence time, and phylogenetic relationships of Allioideae (Amaryllidaceae)[J]. Sci Rep, 11(1): 3262.

KATOH K, STANDLEY DM, 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability[J]. Mol Biol Evol, 30(4): 772-780.

KURTZ JV, CHOUDHURI, OHLEBUSCH E, et al., 2001. REPuter: the manifold applications of repeat analysis on a genomic scale[J]. Nucleic Acids Res, 29(22): 4633-4642.

LI DM, LI J, WANG DR, et al., 2021. Molecular evolution of chloroplast genomes in subfamily Zingiberoideae (Zingiberaceae)[J]. BMC Plant Biol, 21: 1-24.

LI GE, JIANG CJ, QI Y, et al., 2022. Morphological characteristics and identification points of *Camellia minima* and *Camellia cucphuongensis* from Viet-

nam[J]. Southern Hortic, 33(1): 54-60.

LI L, HU YF, HE M, et al., 2021. Comparative chloroplast genomics: insights into the evolution of the chloroplast genome of *Camellia sinensis* and the phylogeny of *Camellia*[J]. BMC Genomics, 22(138).

LI Q, LUO RJ, GE R, et al., 2022. Analysis on codon usage bias of chloroplast genome in *Ampelopsis grossedentata*[J]. Guangdong Agr Sci, 49(11): 162-169.

LI W, ZHANG CP, GUO XP, et al., 2019. Complete chloroplast genome of *Camellia japonica*: genome structures, comparative and phylogenetic analysis[J]. PLoS ONE, 14(5): e0216645.

LI XP, MENG J, ZHANG N, et al., 2018. Comparative analysis of chloroplast genomes of *Aconitum vilmorinianum* and *Aconitum vilmorinianum* var. *patentipilum*[J]. J Chin Med Mat, 41(8): 1812-1820.

LIU H, HUANG Y, DU X, et al., 2012. Patterns of synonymous codon usage bias in the model grass *Brachypodium distachyon*[J]. Genet Mol Res, 11(4): 4695-4706.

LIU Q, LI Y, YANG RM, et al., 2021. Yellow *Camellia*: Resource status and research progress in modern studies[J]. Mod Chin Med, 23(04): 727-733.

LIU S, NI Y, LI J, et al., 2023. CPGView: A package for visualizing detailed chloroplast genome structures[J]. Mol Eco Resourc, 23(3): 694-704.

MINH BQ, SCHMIDT HA, CHERNOMOR O, et al., 2021. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era[J]. Mol Biol Evol, 37(5): 1530-1534.

National Forestry and Grassland Administration of PRC, 2021. List of National Key Protected Wild Plants [EB]. 2021(15).

PAN S, MOU C, WU H, et al., 2020. Phylogenetic and codon usage analysis of atypical porcine pestivirus (APPV)[J]. Virulence, 11(1): 916-926.

PASCAL L, STEPHAN G, 2018. GB2sequin—A file converter preparing custom GenBank files for database submission[J]. Genomics, 111(4): 759-761.

QI YY, XU WJ, XING T, et al., 2015. Synonymous codon usage bias in the plastid genome is unrelated to gene structure and shows evolutionary heterogeneity[J]. Evol Bio Online, 11: 239-253.

SAI X, 2018. Anti-lung cancer effect of *Camellia euphlbia* flowers extract and its potential mechanism of action[D]. Dalian: Dalian University of Technology.

SHI L, CHEN H, JIANG M, et al., 2019. CPGAVAS2, an integrated plastome sequence annotator and analyzer[J]. Nucleic Acids Res, 7(W1): W65-W73.

SHI FC, 2023. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp[J]. iMeta, 2: e107.

SOPHIARANI Y, ARIF U, SUPRIYO C, 2019. Deciphering codon usage patterns and evolutionary forces in chloroplast genes of *Camellia sinensis* var. *assamica* and *Camellia sinensis* var. *sinensis* in comparison to *Camellia pubicosta*[J]. J Inte Agric, 18(12): 2760-2771.

SUEOKA N, 1988. Directional mutation pressure and neutral molecular evolution[J]. Proc Natl Acad Sci USA, 85(8): 2653-2657.

WEI L, HE J, JIA X, et al., 2014. Analysis of codon usage bias of mitochondrial genome in *Bombyx mori* and its relation to evolution[J]. BMC Evol Biol, 14: 262.

WEI SJ, LIUFU YQ, ZHENG HE, et al., 2022. Using phylogenomics to untangle the taxonomic incongruence of yellow-flowered *Camellia* species (Theaceae) in China[J]. J Sys Evol, 00(00): 1-16.

WU LJ, ZHENG HC, CHEN WR, et al., 2020. Performance and thinking on introduction and cultivation of *Camellia* Sect. *Chrysantha* Chang in Fujian[J]. Fujian For Sci Technol, 47(2): 109-115.

WU CS, CHAW SM, HUANG YY, 2013. Chloroplast phylogenomics indicates that *Ginkgo biloba* is sister to cycads[J]. Gen Bio and Evol, 5(2): 243-254.

XIANG H, ZHANG RZ, BUTLER RR, et al., 2015. Comparative analysis of codon usage bias patterns in Microsporidian genomes[J]. PLoS ONE, 10: e0129223.

XIN YX, LI RZ, LI X, et al., 2021. Analysis on codon usage bias of chloroplast genome in *Mangifera indica*[J]. J Centr S Univ For Technol, 41(9): 145-157.

YANG LC, DENG SX, ZHU YQ, et al., 2023. Comparative chloroplast genomics of 34 species in subtribe Swertiinae (Gentianaceae) with implications for its phylogeny[J]. BMC Plant Bio, 23(164): 0-20.

YANG XY, CAI YB, TAN QL et al., 2021. Analysis of codon usage bias in the chloroplast genome of *Ananas comosus*[J]. J Trop Crop, 43(3): 439-446.

ZHANG L, NI S, LI JY, et al., 2019. Analysis of petal nutrition and bioactive components in different periods of *Camellia nitidissima*[J]. For Res, 32(2): 32-38.

ZHANG JW, 2019. The complete chloroplast genome and phylogenetic analysis of endangered species *Syringa pinnatifolia* (Oleaceae)[D]. Xianyang: Northwest A&F University.

ZHANG TW, FANG YJ, WANG XM, et al., 2012. The complete chloroplast and mitochondrial genome sequences of *Boea hygrometrica*: Insights into the evolution of plant organellar genomes[J]. PLoS ONE, 7(1): e30531.

ZHANG W, ZHAO Y, YANG G, et al., 2019. Determination of the evolutionary pressure on *Camellia oleifera* on Hainan Island using the complete chloroplast genome sequence[J]. PeerJ, 7: e7210.

ZHANG Y, NIE X, JIA X, et al., 2012. Analysis of codon usage patterns of the chloroplast genomes in the Poaceae family[J]. Aust J Bot, 60: 461-470.

ZHAO YC, ZHENG H, XU AX, et al., 2016. Analysis of codon usage bias of envelope glycoprotein genes in nuclear polyhedrosis virus (NPV) and its relation to evolution[J]. BMC Genomics, 17: 677-677.

ZHU WY, LONG Y, ZHENG S, et al., 2022. Chloroplast genome structure and characterization of *Melaleuca bracteata*[J]. Mol Plant Breed, 1-14.

Note: Figure translations are in progress. See original paper for figures.