Chloroplast Genome Characteristics and Phylogenetic Analysis of Sycopsis triplinervia Postprint

Authors: Xiong Shuang, Zhou Fuqin, Wang Shidong, Li Rui, Wang Shubao, Huang Yuan

Date: 2025-03-19T00:00:00+00:00

Abstract

Sycopsis triplinervia is an evergreen shrub of the genus Sycopsis in the family Hamamelidaceae. Due to the controversial phylogenetic status of the genus Sycopsis and unclear relationships with closely related genera such as Distyliopsis and Distylium, this study sequenced and assembled the chloroplast genome of S. triplinervia, and conducted comparative genomic analysis and chloroplast genome-based phylogenetic analysis using chloroplast genome data of other Hamamelidaceae species from public databases. The results showed: (1) The genome size was 159,375 bp, encoding 133 genes including 8 rRNA genes, 37 tRNA genes, 87 protein-coding genes, and 1 pseudogene. (2) Thirty-three dispersed repeat sequences, 39 tandem repeat sequences, and 82 simple sequence repeats (SSRs) were detected. (3) Codon usage bias favored A/U-ending codons, with 9 optimal codons, and was primarily influenced by natural selection. (4) The chloroplast genome was relatively conserved compared with closely related species; fifteen highly variable regions identified in Sycopsis have potential value for molecular identification. (5) Phylogenetic analysis revealed Hamamelidaceae as a monophyletic group, with six genera—Hamamelis, Parrotiopsis, Sycopsis, Distyliopsis, Parrotia, and Distylium—forming a strongly supported monophyletic clade with close relationships among them, wherein S. triplinervia was sister to other taxa in this clade. However, Sycopsis, Parrotia, Distyliopsis, and Distylium were all non-monophyletic. This study provides important fundamental data and reference for phylogenetic research of Hamamelidaceae.

Full Text

Analysis of Chloroplast Genomic Characteristics and Phylogeny in *Sycopsis triplinervia*

XIONG Shuang, ZHOU Fuqin, WANG Shidong, LI Rui, WANG Shubao, HUANG Yuan*

School of Life Sciences, Yunnan Normal University, Kunming 650500, China

Abstract: *Sycopsis triplinervia* is an evergreen shrub in the genus *Sycopsis* (Hamamelidaceae). The phylogenetic position of *Sycopsis* remains controversial, and its evolutionary relationships with closely related genera such as *Distyliopsis* and *Distylium* are unclear. Here we sequenced and assembled the chloroplast genome of *S. triplinervia* and conducted comparative genomic and phylogenomic analyses using publicly available chloroplast genomes from other Hamamelidaceae species. The results were as follows: (1) The chloroplast genome of *S. triplinervia* was 159,375 bp in length and encoded 133 genes, including 8 rRNA genes, 37 tRNA genes, 87 protein-coding genes, and 1 pseudogene. (2) A total of 33 dispersed repeats, 39 tandem repeats, and 82 simple sequence repeats (SSRs) were identified. (3) Codon usage was biased toward A/U endings, with nine optimal codons identified, and natural selection was the primary driver of codon usage bias. (4) The chloroplast genome of *S. triplinervia* was highly conserved compared to its close relatives. Fifteen highly variable regions identified in *Sycopsis* have potential value for molecular identification. (5) Phylogenetic analyses indicated that Hamamelidaceae is monophyletic, and six genera—*Hamamelis*, *Parrotiopsis*, *Sycopsis*, *Distyliopsis*, *Parrotia*, and *Distylium*—formed a strongly supported monophyletic clade. Within this clade, *S. triplinervia* was sister to the remaining taxa. However, *Sycopsis*, *Parrotia*, *Distyliopsis*, and *Distylium* were each non-monophyletic. This study provides fundamental data and a valuable reference for further phylogenetic research on Hamamelidaceae.

Keywords: *Sycopsis triplinervia*, chloroplast genome, *Sycopsis*, Hamamelidaceae, phylogenomics

Hamamelidaceae comprises primarily evergreen or deciduous trees and shrubs, including 27 genera and 80–120 species distributed mainly in eastern Asia, with approximately two-thirds of the species concentrated in southern China (Zhang et al., 2003; APG IV, 2016). As a relatively primitive group of angiosperms, Hamamelidaceae fossils have been discovered from the Cretaceous to the Tertiary strata (Zhang et al., 2003). The primitive and complex nature of Hamamelidaceae plants, along with strong differentiation in external characteristics and pollen morphology, make this family an important group for botanists to explore the origin and early diversification of angiosperms (张志耘, 1999).

Sycopsis is a small genus in Hamamelidaceae, mainly distributed in southwestern Chinese provinces and India (Zhang et al., 2003). Since R. Brown established Hamamelidaceae in 1818, researchers have conducted extensive studies on the

family from morphological similarity, trait evolution, and molecular systematics perspectives, establishing at least 14 classification systems (Harms, 1930; Bogle & Philbrick, 1980; Endress, 1989; Li, 1997; 张志耘, 1999; Zhang et al., 2003; APG IV, 2016). Different classification systems have treated *Sycopsis*, *Distyliopsis*, *Parrotia*, *Distylium*, and *Shaniodendron* inconsistently. Based on morphological characteristics, Endress (1970) separated four species originally belonging to *Sycopsis* to establish the new genus *Distyliopsis*, considering it more distantly related to *Sycopsis* but closely related to *Distylium*. The *Flora Reipublicae Popularis Sinicae* revised the classification system for Chinese Hamamelidaceae, recognizing nine species in *Sycopsis* divided into two subgenera, and did not support the separation of *Distyliopsis* from *Sycopsis* (张宏达, 1979). Combining morphological and molecular systematic evidence, Li (1997) and *Flora of China* recognized only 2–3 species in *Sycopsis*, with both *Sycopsis sinensis* and *S. triplinervia* being endemic to China, suggesting that separating *Distyliopsis* from *Sycopsis* is reasonable (Zhang et al., 2003). The APG IV system (2016) also supports the independence of *Distyliopsis*, transferring five species originally belonging to the subgenus *Metasycopsis* to *Distyliopsis*. *Sycopsis*, *Distyliopsis*, *Parrotia*, and *Distylium* can hybridize with each other, with hybrid characteristics often intermediate between the two parents (Endress, 1989; Johnson, 2024). Therefore, classification of *Sycopsis* species based solely on morphological characteristics is often inaccurate, and species delimitation should integrate multiple criteria including morphological characteristics, population genetic differentiation, and phylogenetic relationships (刘建全, 2016). In a phylogenetic tree based on the chloroplast gene *matK*, *Sycopsis* and *Distylium* formed sister groups (Li et al., 1999a). Meanwhile, nuclear ITS sequence studies found that *Distylium* and *Distyliopsis* formed a sister clade, while *Sycopsis* was more closely related to *Parrotia* and *Shaniodendron* (Li et al., 1999b). Wang et al. (2022) constructed a phylogenetic tree of Hamamelidaceae based on complete chloroplast genomes and found that *Sycopsis sinensis* was nested within *Distylium*. In these previous studies, only one species from *Sycopsis* was sampled, and some lacked *Distyliopsis*, failing to clarify whether *Sycopsis* is monophyletic and its systematic relationships with closely related genera. Therefore, broader sampling of *Sycopsis* and additional molecular systematic evidence are needed to resolve the phylogenetic relationships among *Sycopsis*, *Distyliopsis*, *Distylium*, and other related genera.

Although molecular systematic studies based on fragment sequences such as *rbcL*, *matK*, and ITS can reveal phylogenetic relationships at different taxonomic levels, the limited information sites and different evolutionary rates of these fragments often result in inconsistent phylogenetic trees (陈丽琼等, 2022). To construct more reliable phylogenetic trees, integration of more gene or genomic data is necessary. With the development of next-generation sequencing technology, complete chloroplast genomes have been widely used to study plant phylogeny. Chloroplast genomes are typically 107–218 kb in size, with low sequencing costs, and offer advantages over nuclear genomes including conservation, genetic stability, and absence of genetic recombination, making them more

suitable for plant phylogenetic and evolutionary studies (Corriveau & Coleman, 1988; Grevich & Daniell, 2005; Ravi et al., 2008; Daniell et al., 2016). The chloroplast genome of *Sycopsis sinensis* has been reported in detail (Peng et al., 2020), but chloroplast genome characteristics and phylogenetic analyses for other species in *Sycopsis* have not been reported.

Therefore, this study sequenced, assembled, and annotated the chloroplast genome of *Sycopsis triplinervia*, conducted comparative genomic and phylogenetic analyses using publicly available Hamamelidaceae chloroplast genome data, and addressed the following scientific questions: (1) What is the molecular structure of the *S. triplinervia* chloroplast genome? (2) What are the repeat sequences, SSR loci, and codon usage bias in the *S. triplinervia* chloroplast genome? (3) Can we construct a Hamamelidaceae phylogenetic tree based on chloroplast genome data to analyze the systematic position of *S. triplinervia* and provide new molecular evidence for exploring relationships among *Sycopsis*, *Distyliopsis*, *Parrotia*, and *Distylium*?

1.1 Plant Material, DNA Extraction, and Sequencing

Plant material of *Sycopsis triplinervia* was collected from Sanjiangkou Forest Farm, Daguan County, Zhaotong City (103°56 20 E, 28°12 47 N). Fresh leaves from healthy plants were dried with silica gel, assigned sample number 08CS358, and the voucher specimen was deposited at the Herbarium of Kunming Institute of Botany, Chinese Academy of Sciences (accession: KUN1325573). DNA was extracted from dried leaves using a modified CTAB method (Porebski et al., 1997), and DNA purity was analyzed (OD_{260}/OD_{280} = 1.91; concentration = 160.96 ng · L^{-1}). After quality assessment, the extracted DNA was fragmented to 150 bp by sonication for library construction and high-throughput sequencing, yielding 2.1 Gb of raw data. The raw sequencing data have been uploaded to the NCBI database (Submission ID: SUB14370161; BioProject ID: PRJNA1098481).

1.2 Chloroplast Genome Assembly and Annotation

Raw data were quality-controlled using fastp v0.23.2 (Chen, 2023), and the chloroplast genome was assembled using GetOrganelle v1.7.3.5 (Jin et al., 2020). The assembled sequence was imported into Geneious Prime 2023 to identify inverted repeats using Repeat Finder, and the highest similarity sequence [*Distylium racemosum*] was found via NCBI BLAST as a reference. Annotation was performed using the online tool GeSeq (https://chlorobox.mpimp-golm.mpg.de/geseq.html) (Tillich et al., 2017). The annotated sequence was manually corrected in Geneious Prime 2023 and uploaded to NCBI (accession: PP625971). Finally, the circular physical map of the *S. triplinervia* chloroplast genome was drawn using the online software CPGView (http://www.1kmpg.cn//cpgview).

1.3 Analysis of Chloroplast Genome Repeats and SSRs

Dispersed repeats in the *S. triplinervia* chloroplast genome were detected using REPuter (Kurtz et al., 2001) with the following parameters: minimum repeat length of 30 bp and Hamming distance of 3. Simple sequence repeats (SSRs) were analyzed using MISA (Beier et al., 2017) with minimum repeat thresholds of 10, 5, 4, 3, 3, and 3 for mono- to hexanucleotides, respectively. Tandem repeats were identified using Tandem Repeats Finder (Benson, 1999).

1.4 Codon Usage Bias Analysis

Protein-coding sequences (CDS) were extracted from the *S. triplinervia* chloroplast genome using Geneious Prime 2023. CDS sequences $ $300 bp were selected, with only one copy retained for duplicate sequences, ensuring that sequences contained only A, T, C, G bases and no internal stop codons. CodonW v1.4.2 (Peden, 2005) was used to calculate relative synonymous codon usage (RSCU) and effective number of codons (ENC). The EMBOSS CUSP program (https://www.bioinformatics.nl/emboss-explorer/) calculated total GC content and GC content at the three codon positions (GCall, GC1, GC2, GC3) for each CDS. The average of GC1 and GC2 and the GC content at synonymous third positions were recorded as GC12 and GC3S, respectively.

ENC values were sorted to establish high- and low-expression gene libraries by selecting five genes from each end. ΔRSCU values were calculated, and codons with $\Delta\text{RSCU} > 0.08$ were designated as high-expression codons. Optimal codons were identified by combining high-frequency codons ($\text{RSCU} > 1$) with high-expression codons ($\Delta\text{RSCU} > 0.08$). ENC-plot scatter diagrams were generated with ENC on the y-axis and GC3S on the x-axis. Neutrality plots were created with GC12 on the y-axis and GC3 on the x-axis, including fitted regression lines and correlation calculations. PR2-plot analysis was performed using MEGA v7.0 to calculate third-position base composition (A3, T3, C3, G3), plotting A3/(A3+T3) against G3/(G3+C3) for each CDS. All visualizations were completed using R package ggplot2 v3.4.2.

1.5 Comparative Analysis of Chloroplast Genomes

IR boundary information was compared for seven Hamamelidaceae chloroplast genome sequences using the online tool CPJS draw (Li et al., 2023): *Sycopsis triplinervia* (PP625971), *Sycopsis sinensis* (NC071198), *Distyliopsis dunnii* (NC071205), *Distyliopsis laurifolia* (NC071202), *Distylium buxifolium* (NC059888), *Distylium chinense* (NC059885), and *Parrotia persica* (NC071840). Collinearity among these seven sequences was analyzed using the Mauve plugin in Geneious Prime 2023 (Darling et al., 2004). Using *Distylium* as the reference, sequence identity across the seven chloroplast genomes was analyzed via mVISTA (Frazer et al., 2004) in LAGAN mode.

1.6 Nucleotide Polymorphism Analysis

DnaSP v6.0 (Rozas et al., 2017) was used to analyze nucleotide polymorphism (Pi) in protein-coding and non-coding regions (including introns and intergenic spacers) of four *Sycopsis* chloroplast genomes: *Sycopsis triplinervia* PP625971, *Sycopsis sinensis* NC071198, *Sycopsis sinensis* MT323104, and *Sycopsis sinensis* MN496080. Coding regions were extracted using Geneious Prime 2023. Non-coding regions were extracted using a Perl script (3_{{{extract}}*{{bed}}*{{{coding}}*{{and}}*{noncoding}) and manually verified. Extracted sequences were aligned with MAFFT and imported into DnaSP v6.0 for polymorphism analysis.

1.7 Phylogenetic Analysis

To clarify the phylogenetic position of *S. triplinervia* within Hamamelidaceae, 60 chloroplast genome sequences from 53 Hamamelidaceae species and four chloroplast genome sequences from three Altingiaceae species were downloaded from NCBI. Combined with the newly assembled *S. triplinervia* genome, 65 sequences were used for phylogenetic analysis. Using Altingiaceae as the outgroup, both a shared CDS phylogenetic tree and a complete chloroplast genome phylogenetic tree were constructed.

For the shared CDS tree, *Distylium* and *Sycopsis sinensis* were used as reference genomes for batch annotation of all 65 chloroplast genomes using GeSeq (Tillich et al., 2017). A Python script extracted 78 shared CDS, which were individually aligned using MAFFT v7.525 (Katoh et al., 2002). IQ-TREE v2.1.4 (Nguyen et al., 2015) was used to construct the phylogenetic tree.

For the complete genome tree, all 65 sequences were aligned using MAFFT v7.525 to create a supermatrix. Trimal v1.4 (Capella-Gutiérrez et al., 2009) removed poorly aligned regions with parameters "gt = 0.75, resoverlap = 0.90, seqoverlap = 0.90" . IQ-TREE v2.1.4 constructed the phylogeny using the best-fit TVM+F+I+R3 substitution model. Both phylogenetic trees were evaluated with 10,000 SH-aLRT and ultrafast bootstrap (UFboot) replicates (Guindon et al., 2010; Hoang et al., 2018) and visualized in FigTree.

2.1 Chloroplast Genome Structure and Basic Features of *Sycopsis triplinervia*

The complete chloroplast genome of *S. triplinervia* was 159,375 bp with a total GC content of 38% [Figure 1: see original paper]. The large single-copy (LSC) region was 88,067 bp (36.2% GC), the small single-copy (SSC) region was 18,808 bp (32.5% GC), and each inverted repeat (IR) was 26,250 bp (43.1% GC), substantially higher than the LSC and SSC regions . The genome encoded 133 genes, including 8 rRNA genes, 37 tRNA genes, 87 protein-coding genes, and 1 pseudogene (*#ycf1*) . Genes were functionally categorized into four main groups: photosynthesis-related, self-replication, other genes, and un-

known function genes. Seventeen genes had two copies (*ndhB*, *rpl2*, *rrn4.5*, *trnA-UGC*, *ycf15*, etc.), while the rest were single-copy. Eighteen genes contained introns: 15 with one intron (*atpF*, *ndhA*, *ndhB*, *petB*, *petD*, *rpl16*, *rpl2*, *rps16*, *rpoC1*, *trnA-UGC*, *trnG-GCC*, *trnI-GAU*, *trnK-UUU*, *trnL-UAA*, *trnV-UAC*) and three with two introns (*rps12*, *clpP*, *ycf3*).

2.2 Repeat Sequences and SSR Loci Analysis

REPuter identified 33 dispersed repeats (30–69 bp) in the *S. triplinervia* chloroplast genome, comprising 1 forward repeat, 17 palindromic repeats, and 15 reverse repeats; no complementary repeats were detected. Twenty-two repeats (14 forward, 7 palindromic, 1 reverse) were located in genes (*ycf2*, *psaB*, *psaA*, *ycf3*, *ndhA*, *rps12*, *rps3*, *ccsA*, *ndhD*); three palindromic repeats spanned both genes (*ycf3* and *ndhA*) and intergenic spacers (IGS); four repeats (1 forward, 3 palindromic) were in IGS regions; and four (2 forward, 2 palindromic) were in introns. Tandem Repeats Finder predicted 39 tandem repeats ranging from 24 to 132 bp, predominantly distributed in IGS regions (64.10%).

SSR analysis identified 82 SSR loci [Figure 2: see original paper], including 61 mononucleotides, 9 dinucleotides, 3 trinucleotides, 7 tetranucleotides, and 2 pentanucleotides. Most SSRs (70.73%) were located in IGS regions. A/T bases occurred most frequently in SSRs, with 26 mononucleotide SSRs containing A and 33 containing T.

2.3 Codon Usage Bias Analysis

Analysis of 53 protein-coding genes revealed 21,394 codons . Leucine was most abundant (2,221 codons, 10.38%), while cysteine was least abundant (243 codons, 1.14%). Among all codons, AUU (isoleucine) was most frequent (885 occurrences), whereas UGC (cysteine) was least frequent (54). Among stop codons, UAA was most frequent (50.94%), followed by UAG (24.53%) and UGA (24.53%). The highest RSCU value was for GCU (alanine, 1.88) and the lowest for AGC (serine, 0.33). Thirty codons had RSCU $>$ 1, and except for UUG (leucine), all ended with A/U, indicating a preference for A/U-ending codons.

Based on ENC values, high- and low-expression gene libraries were constructed, yielding 28 high-expression codons (ΔRSCU $>$ 0.08), 3/7 of which ended with A/U. Nine optimal codons were identified by combining high-frequency (RSCU $>$ 1) and high-expression codons , all ending with A/U.

PR2-plot analysis showed that most points fell in the region where A3/(A3+T3) $<$ 0.5 and G3/(G3+C3) $>$ 0.5, indicating higher usage of G and T than C and A at the third codon position, suggesting both mutation and natural selection influence codon bias [Figure 3A: see original paper]. ENC-plot analysis showed most genes fell below the expected curve, indicating weak mutational effects [Figure 3B: see original paper]. Neutrality plot analysis revealed a negative but non-significant correlation between GC12 and GC3 (R = -0.034, slope =

-0.0494), suggesting natural selection is the dominant force shaping codon usage bias [Figure 3C: see original paper].

2.4 Comparative Analysis of Chloroplast Genomes

IR boundary analysis of *S. triplinervia* and six related species revealed conserved IR lengths (26,218–26,258 bp) [Figure 4: see original paper]. Except for *S. triplinervia*, *Distyliopsis laurifolia*, and *Parrotia persica*, the LSC/IRB boundary was within the *rps19* gene for the other four species. The *ndhF* gene spanned the IRB/SSC boundary in *D. laurifolia* but was 10–16 bp from the boundary in the other six species. The SSC/IRA boundary was located within *ycf1* in all seven species, with 1,017 bp of *ycf1* in IRA. In *S. triplinervia* and *D. laurifolia*, *ycf1* extended 4,581 bp and 4,554 bp into the SSC region, respectively, while the other five species extended 4,560 bp. The *trnH* gene was farthest from the IRA/LSC boundary in *S. triplinervia* (122 bp), while the distance varied only 7–25 bp in the other six species.

Mauve analysis showed high collinearity among the seven chloroplast genomes, with no large-scale gene rearrangements or inversions, indicating a conserved structure [Figure 5: see original paper]. mVISTA analysis using *Distylium* as reference revealed that coding and IR regions were more conserved than non-coding regions [Figure 6: see original paper]. Overall, the seven species showed high similarity with no large fragment deletions.

2.5 Nucleotide Polymorphism Analysis

To identify highly variable regions in *Sycopsis*, DnaSP v6.0 analyzed nucleotide polymorphism in protein-coding and non-coding regions of four chloroplast genomes. Among 79 protein-coding regions (counting duplicate genes once), *psaJ* showed the highest Pi value (0.00741), followed by *petL* (0.00521) [Figure 7A: see original paper]. Thirty-five highly conserved regions (Pi = 0) were identified. Using Pi > 0.003 as the threshold, seven highly variable genes were identified: *psaJ* (0.00741), *petL* (0.00521), *psaI* (0.0045), *rpl16* (0.00449), *petG* (0.00439), *rps18* (0.00327), and *rpl32* (0.00303), six of which were in the LSC region.

Non-coding region analysis showed the highest Pi in *rps2-rpoC2* (0.01458), followed by *rps14-psaB* (0.01378) [Figure 7B: see original paper]. Sixty-four highly conserved non-coding regions (Pi = 0) were found. Using Pi > 0.006 as the threshold, eight highly variable fragments were identified: *rps2-rpoC2* (0.01458), *rps14-psaB* (0.01378), *ndhD-psaC* (0.00909), *psbL-psbF* (0.00862), *accD-psaI* (0.00722), *trnW-CCA-trnP-UGG* (0.00671), *ndhA* intron (0.00639), and *rpl2__2-trnH-GUG* (0.00617), six of which were in the LSC region. Mean Pi values were 0.00093 for coding regions and 0.00147 for non-coding regions, with IR regions showing significantly lower polymorphism than LSC and SSC regions.

2.6 Phylogenetic Analysis

Maximum likelihood (ML) trees were constructed using 61 chloroplast genomes from 54 Hamamelidaceae species plus four from three Altingiaceae outgroup species [FIGURE:8, FIGURE:9]. Both trees strongly supported Hamamelidaceae as monophyletic (SH-aLRT/UFboot = 100/100). Among 16 Hamamelidaceae genera, *Rhodoleia* and *Exbucklandia* formed one clade (100/100), sister to the remaining genera (100/100). *Mytilaria* and *Chunia*, *Disanthus*, *Loropetalum* and *Corylopsis*, *Sinowilsonia*, and *Eustigma* and *Fortunearia* each formed monophyletic clades (100/100). Six genera—*Hamamelis*, *Parrotiopsis*, *Sycopsis*, *Distyliopsis*, *Parrotia*, and *Distylium*—formed a strongly supported monophyletic clade (Clade 1, 100/100), with *S. triplinervia* sister to other members. However, *Sycopsis*, *Parrotia*, *Distyliopsis*, and *Distylium* were each non-monophyletic.

Topologies of the complete genome and shared CDS trees were largely consistent, with minor differences in the placement of *Corylopsis veitchiana* and *Distylium cuspidatum*. Both trees showed that three individuals of *Sycopsis sinensis* did not form a monophyletic group, with one individual (NC071189) sister to *Distylium chinense*.

3.1 Conservative Chloroplast Genome Structure

The *S. triplinervia* chloroplast genome is highly conserved in structure, size, gene number, total GC content, and regional length and GC content, consistent with other Hamamelidaceae species in *Loropetalum*, *Corylopsis*, and *Distylium* (Dong et al., 2021; Wang et al., 2022). Total GC content was lower than AT content, with IR regions showing higher GC content than LSC and SSC regions. High AT content may result from synonymous codons ending predominantly in A/U, related to natural selection and mutation during evolution (Shimda & Sugiuro, 1991; Clegg et al., 1994; Liu et al., 2019). Additionally, rRNA genes located in IR regions have high GC content, contributing to the relatively higher GC content of IR regions (Zhang et al., 2012; Wang et al., 2022).

The *S. triplinervia* chloroplast genome encodes 133 genes, including the pseudogene *#ycf1*, similar to *Loropetalum chinense* and *Corylopsis velutina* (Wang et al., 2022). Various mutation events during chloroplast genome evolution, such as insertions-deletions (InDels), substitutions, inversions, and copy number variations, can cause gene loss or pseudogenization (Bendich et al., 1987; Kumar et al., 2014; Abdullah et al., 2021). Pseudogenization is the process by which a functional gene becomes non-functional, and pseudogenes often share high homology with functional genes (Wickett et al., 2011; Van et al., 2014). The pseudogene *#ycf1* in *S. triplinervia* is 4,557 bp shorter than the functional *ycf1* gene, suggesting pseudogenization was primarily caused by base deletion.

3.2 Evolutionary Characteristics of the *Sycopsis triplinervia* Chloroplast Genome

Repeat sequences play important roles in analyzing base substitution, genome rearrangement, and phylogeny (Nie et al., 2012). SSRs in the *S. triplinervia* chloroplast genome are predominantly mononucleotide A/T repeats, consistent with other Hamamelidaceae species (Wang et al., 2022). This further validates that chloroplast SSRs mainly consist of A/T repeats rather than C/G repeats (Kuang et al., 2011). Chloroplast genomes typically have high AT content (Asaf et al., 2018; Liu et al., 2019), and high A/T content in SSRs may contribute to the overall AT bias.

Codon usage bias is an important evolutionary feature in both prokaryotes and eukaryotes (Sharp et al., 1988; Wang et al., 2011). The *S. triplinervia* chloroplast genome prefers UAA as the stop codon and favors A/U-ending codons, with all nine optimal codons ending in A/U, likely related to high AT content. Codon bias is influenced primarily by mutation and natural selection (Sharp et al., 2010). Neutrality plot and ENC-plot analyses indicated natural selection is the dominant force shaping codon bias in *S. triplinervia*, consistent with studies on Juglandaceae and Asteraceae but contrasting with *Arabidopsis* and maize where mutation is the primary factor (Zhou et al., 2008; Nie et al., 2014; Zeng et al., 2023). This suggests that codon preferences are influenced by multiple factors that vary among plant lineages.

The lack of recombination and presence of IR regions allow chloroplast genomes to better retain evolutionary history. IR expansion or contraction is a major cause of size variation, and IR expansion can place IR/SC boundaries within coding regions, creating pseudogenes (Ravi et al., 2008). In *S. triplinervia*, the IRA/SSC boundary lies within *ycf1*, creating the pseudogene #*ycf1* at the IRB/SSC boundary, possibly resulting from gene loss events during Hamamelidaceae evolution (Wang et al., 2022). The 15 highly variable regions identified (e.g., *psaJ*, *petL*, *rps14-psaB*) can serve as candidate molecular markers for investigating *Sycopsis* phylogeny.

3.3 Systematic Position and Evolutionary Relationships of *Sycopsis*

Chloroplast genomes are widely used in plant phylogenetic analysis due to their conserved structure, moderate evolutionary rate, and ease of sequencing (Wolfe et al., 1987; Grevich & Daniell, 2005; Daniell et al., 2016). This study reconstructed intergeneric relationships in Hamamelidaceae using ML analysis of chloroplast genomes, revealing that *Hamamelis*, *Parrotiopsis*, *Sycopsis*, *Distyliopsis*, *Parrotia*, and *Distylium* form a strongly supported monophyletic clade. Early morphological frameworks suggested close relationships between *Sycopsis* and *Distylium* (Walker, 1944). Endress (1970) established *Distyliopsis* based on petal presence/absence and capsule persistence, considering it more closely related to *Distylium*. Molecular studies using *matK* and ITS also supported close

relationships among these genera and the separation of *Distyliopsis* from *Sycopsis* (Li et al., 1999a, b; Xiang et al., 2019). Morphologically, *Sycopsis*, *Distyliopsis*, and *Distylium* differ in inflorescence, venation, and capsule arrangement. Our study found species from *Sycopsis* and *Distyliopsis* nested within *Distylium*, with none of these three genera forming monophyletic groups, supporting the establishment of *Distyliopsis* but requiring deeper investigation of species relationships. However, Dong et al. (2021) supported *Distylium* monophyly based on chloroplast genomes, but their study lacked *Sycopsis* and *Distyliopsis* species, and incomplete sampling may have affected conclusions.

Our results show that two *Sycopsis* species and three individuals of *S. sinensis* failed to form monophyletic groups, possibly due to large genetic differentiation among populations, hybrid origin of some individuals, or the conservative nature of chloroplast genome sequences. *Sycopsis triplinervia* showed close relationships with *Distyliopsis laurifolia* and *Parrotia persica*, possibly related to shared characteristics of lacking petals and having stellate tomentum on young shoots (Zhang et al., 2003). While this study clarifies relationships among *Sycopsis*, *Distyliopsis*, *Parrotia*, and *Distylium* to some extent, resolving the systematic relationships of *Sycopsis* and its relatives requires more comprehensive sampling and additional nuclear DNA information from high-throughput sequencing technologies (e.g., whole-genome resequencing, transcriptome sequencing, genome skimming).

Acknowledgments

We thank the Germplasm Bank of Wild Species in Southwest China and the National Wild Plant Germplasm Resource Center for providing *Sycopsis triplinervia* material and for their strong support of this research.

References

ABDULLAH, MEHMOOD F, HEIDARI P, et al., 2021. Pseudogenization of the chloroplast threonine (trnT-GGU) gene in the sunflower family (Asteraceae) [J]. Scientific Reports, 11(1).

APG IV, 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV [J]. Botanical Journal of the Linnean Society, 181(1).

ASAF S, KHAN AL, KHAN MA, et al., 2018. Complete chloroplast genome sequence and comparative analysis of loblolly pine (*Pinus taeda* L.) with related species [J]. PLoS One, 13(3): e0192966.

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

BENDICH AJ, 1987. Why do chloroplasts and mitochondria contain so many copies of their genome? [J]. Bioessays, 6(6): 279-282.

BENSON G, 1999. Tandem repeats finder: a program to analyze DNA sequences [J]. Nucleic Acids Research, 27(2): 573-580.

BOGLE AL, PHILBRICK CT, 1980. A generic atlas of Hamamelidaceous pollens [J]. Contributions from the Gray Herbarium of Harvard University, 210: 29-103.

CAPELLA-GUTIÉRREZ S, SILLA-MARTÍNEZ JM, GABALDÓN T, 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses [J]. Bioinformatics, 25(15).

CHEN LQ, ZHANG ZR, YANG JB, et al., 2022. Plastid phylogenomic insights into the phylogeny of Convolvulaceae [J]. Guihaia, 42(10):1740-1749.

CHEN SF, 2023. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp [J]. iMeta, 2(2): e107.

CLEGG MT, GAUT BS, LEARN GH, et al., 1994. Rates and patterns of chloroplast DNA evolution [J]. Proceedings of the National Academy of Sciences, 91(15): 6795-6801.

CORRIVEAU JL, COLEMAN AW, 1988. Rapid screening method to detect potential biparental inheritance of plastid DNA and results for over 200 angiosperm species [J]. American Journal of Botany, 75(10): 1443-1458.

DANIELL H, LIN C-S, YU M, et al., 2016. Chloroplast genomes: diversity, evolution, and applications in genetic engineering [J]. Genome Biology, 17(1): 134.

DARLING ACE, MAU B, BLATTNER FR, et al., 2004. Mauve: Multiple alignment of conserved genomic sequence with rearrangements [J]. Genome Research, 14(7): 1394-1403.

DONG W, LIU Y, XU C, et al., 2021. Chloroplast phylogenomic insights into the evolution of *Distylium* (Hamamelidaceae) [J]. BMC Genomics, 22(1): 293.

ENDRESS P K, 1989. A suprageneric taxonomic classification of the Hamamelidaceae [J]. Taxon, 38(3): 371-376.

ENDRESS PK, 1970. Die Infloreszenzen der apetalen Hamamelidaceen, ihre grundsätzliche morphologische und systematische Bedeutung [J]. Botanische Jahrbücher für Systematik, 90.

FRAZER KA, PACHTER L, POLIAKOV A, et al., 2004. VISTA: computational tools for comparative genomics [J]. Nucleic Acids Research, 32(2): W273-W279.

GREVICH JJ, DANIELL H, 2005. Chloroplast genetic engineering: Recent advances and future perspectives [J]. Critical Reviews in Plant Sciences, 24(2): 83-107.

GUINDON S, DUFAYARD JF, LEFORT V, et al., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0 [J]. Systematic Biology, 59(3): 307-321.

HARMS H, 1930. Hamamelidaceae [M]. In: ENGLER A and PRANTL K editors. Die Naturlichen Pflanzenfamilien. Vol. 18a. Leipzig: Verlag von Wilhelm Engelmann: 303-345.

HOANG DT, CHERNOMOR O, VON HAESELER A, et al., 2018. UFBoot2: Improving the ultrafast bootstrap approximation [J]. Molecular Biology and Evolution, 35(2): 518-522.

JIN JJ, YU WB, YANG JB, et al., 2020. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes [J]. Genome Biology, 21(1): 241.

JOHNSON O, 'Distylium' from the website Trees Shrubs Online (treesandshrubsonline.org/articles/distylium/) [EB/OL]. Accessed 2024-08-04.

KATOH K, MISAWA K, KUMA KI, et al., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform [J]. Nucleic Acids Research, 30(14).

KUANG DY, Wu H, WangYL, et al., 2011. Complete chloroplast genome sequence of *Magnolia kwangsiensis* (Magnoliaceae): implication for DNA barcoding and population genetics [J]. Genome, 54(8): 663-673.

KUMAR RA, OLDENBURG DJ, BENDICH AJ, 2014. Changes in DNA damage, molecular integrity, and copy number for plastid DNA and mitochondrial DNA during maize development [J]. Journal of Experimental Botany, 65(22): 6425-6439.

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

LI H, GUO Q, XU L, et al., 2023. CPJSdraw: analysis and visualization of junction sites of chloroplast genomes [J]. PeerJ, 11: e15326.

LI JH, 1997. Systematics of the Hamamelidaceae based on morphological and molecular evidence [D]. Durham: University of New Hampshire: 251-273.

LI JH, BOGLE AL, KLEIN AS, 1999a. Phylogenetic relationships in the Hamamelidaceae: evidence from the nucleotide sequences of the plastid gene *matK* [J]. Plant Systematics and Evolution, 218(3/4): 205-219.

LI JH, BOGLE AL, KLEIN AS, 1999b. Phylogenetic relationships of the Hamamelidaceae inferred from sequences of internal transcribed spacers (ITS) of nuclear ribosomal DNA [J]. American Journal of Botany, 86(7): 1027-1037.

LIU JQ, 2016. "The integrative species concept" and "species on the speciation way" [J]. Biodiversity Science, 24(9): 1004-1008.

LIU X, CHANG E-M, LIU JF, et al., 2019. Complete chloroplast genome sequence and phylogenetic analysis of *Quercus bawanglingensis* Huang, Li et Xing, a Vulnerable Oak Tree in China [J]. Forests, 10(7): 587.

NGUYEN LT, SCHMIDT HA, VON HAESELER A, et al., 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies [J]. Molecular Biology and Evolution, 32(1): 268-274.

NIE X, DENG P, FENG K, et al., 2014. Comparative analysis of codon usage patterns in chloroplast genomes of the Asteraceae family [J]. Plant Molecular Biology Reporter, 32(4).

NIE X, LV S, ZHANG Y, et al., 2012. Complete chloroplast genome sequence of a major invasive species, crofton weed (*Ageratina adenophora*) [J]. PLoS ONE, 7(5): e36869.

PEDEN J, 2005. CodonW version 1.4.2 [CP]. Nottingham, UK: University of Nottingham.

PENG Y, YANG L, WEI J, 2020. The complete chloroplast genome of *Sycopsis sinensis* Oliver [J]. Mitochondrial DNA Part B Resources, 5(3): 3002-3003.

POREBSKI S, BAILEY LG, BAUM BR, 1997. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components [J]. Plant Molecular Biology Reporter, 15(1): 8-15.

RAVI V, KHURANA JP, TYAGI AK, et al., 2008. An update on chloroplast genomes [J]. Plant Systematics and Evolution, 271(1/2): 101-122.

ROZAS J, FERRER-MATA A, CARLOS SANCHEZ-DELBARRIO J, et al., 2017. DnaSP 6: DNA sequence polymorphism analysis of large data sets [J]. Molecular Biology and Evolution, 34(12): 3299-3302.

SHARP PM, COWE E, HIGGINS DG, et al., 1988. Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity [J]. Nucleic Acids Research, 16(17): 8207-8211.

SHARP PM, EMERY LR, ZENG K, 2010. Forces that influence the evolution of codon bias [J]. Philosophical Transactions of the Royal Society B-biological Sciences, 365(1544): 1203-1212.

SHIMDA H, SUGIURO M, 1991. Fine structural features of the chloroplast genome: comparison of the sequenced chloroplast genomes [J]. Nucleic Acids Research, 19(5): 983-995.

TILLICH M, LEHWARK P, PELLIZZER T, et al., 2017. GeSeq-versatile and accurate annotation of organelle genomes [J]. Nucleic Acids Research, 45(W1): W6-W11.

VAN DER BURGT A, JASHNI MK, BAHKALI AH, et al., 2014. Pseudogenization in pathogenic fungi with different host plants and lifestyles might reflect

their evolutionary past [J]. Molecular Plant Pathology, 15(2): 133-144.

WALKER EH, 1944. A revision of *Distylium* and *Sycopsis* (Hamamelidaceae) [J]. Journal of the Arnold Arboretum, 25(3): 319-341.

WANG B, YUAN J, LIU J, et al., 2011. Codon usage bias and determining forces in green plant mitochondrial genomes [J]. Journal of Integrative Plant Biology, 53(4): 324-334.

WANG NJ, CHEN SF, XIE L, et al., 2022. The complete chloroplast genomes of three Hamamelidaceae species: Comparative and phylogenetic analyses [J]. Ecology and Evolution, 12(2): e8637.

WICKETT NJ, FORREST LL, BUDKE JM, et al., 2011. Frequent pseudogenization and loss of the plastid-encoded sulfate-transport gene *cysA* throughout the evolution of liverworts [J]. American Journal of Botany, 98(8): 1263-1275.

WOLFE KH, LI WH, SHARP PM, 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs [J]. Proceedings of the National Academy of Sciences of the United States of America, 84(24): 9054-9058.

XIANG XG, XIANG KL, ORTIZ RDC, et al., 2019. Integrating palaeontological and molecular data uncovers multiple ancient and recent dispersals in the pantropical Hamamelidaceae [J]. Journal of Biogeography, 46(11): 2622-2631.

ZENG YJ, SHEN LW, CHEN SQ, et al., 2023. Codon usage profiling of chloroplast genome in Juglandaceae [J]. Forests, 14(2): 378.

ZHANG HD, 1979. *Flora Reipublicae Popularis Sinicae*: Vol. 35, No. 2 [M]. Beijing: Science Press: 36-116.

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 ZY, 1999. Notes on the modern classification systems of the Hamamelidaceae [J]. Acta Botanica Yunnanica, 21(1): 1-10.

ZHANG ZY, ZHANG HD, ENDRESS PK, 2003. *Flora of China* [M]//WU ZY, RAVEN PH, editors. Vol. 9. Beijing: Science Press; St. Louis: Missouri Botanical Garden Press: 18-42.

ZHOU M, WEI L, XIA L, 2008. Patterns of synonymous codon usage bias in chloroplast genomes of seed plants [J]. Forest Ecosystems, 10(4): 235-242.

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