BMC Plant Biology

RESEARCH

Open Access

Comparative chloroplast genomes of Hamamelidaceae: genome evolution and phylogenomic implications

Yunyan Zhang[1*], Yahong Wang[1], Qixun Chen[1], Zhiyuan Li[1], David Y. P. Tng[2], Shuang Wang[1,3] and Zhongsheng Wang[1*]

Abstract

Background Hamamelidaceae, a family within the core eudicots and the order Saxifragales, represents a characteristic lineage of the East Asian flora and holds considerable ecological and economic importance. However, molecular phylogenetic studies based on single or a few genetic loci have failed to resolve certain phylogenetic relationships among species, genera, and tribes within the family. Besides, less knowledge on the chloroplast genome evolution of Hamamelidaceae has known to us so far.

Results and conclusion In this study, we first sequenced and *de novo* assembled the complete chloroplast genomes of 37 representative species with comprehensive genus-level sampling within Hamamelidaceae. The whole length of chloroplast genomes of the 37 Hamamelidaceae species ranged from 158,090 bp to 160,814 bp with typical quadripartite structure. Minor visible divergences exhibited in the inverted repeat (IR)/single-copy (SC) boundary regions, which leading to the size variations in Hamamelidaceae chloroplast genomes. Four DNA barcodes (*ndhC/rps18/trnT-GGU/trnR-UCU*) and 6,084 chloroplast microsatellites were identified for Hamamelidaceae. Comparative chloroplast genomes analysis revealed that the genome structure, gene arrangement, gene and GC content, codon usage bias, and core genes of Hamamelidaceae species were highly conserved, while *clpP* gene exhibited significant positive selection (dN/dS > 1) in Hamamelidaceae. Chloroplast genome-based phylogeny of Hamamelidaceae supported the monophyly of the majority of extant genera and all Hamamelidaceae species were divided into four subfamilies with high resolution. Exbucklandioideae was the earliest-diverging lineage, followed by Mytilarioideae, Disanthoideae, and Hamamelidoideae. Moreover, chloroplast capture events through hybridization or introgression may lead to the cytonuclear discordance. Overall, our study not only illuminate the chloroplast genome evolution but also provide a robust phylogenetic framework for constructing a more resolved "Tree of Life" of Hamamelidaceae.

Keywords Hamamelidaceae, Comparative chloroplast genomics, Genome evolution, Phylogeny of Hamamelidaceae, Chloroplast genetic resources

*Correspondence:
Yunyan Zhang
zyynju@nju.edu.cn
Zhongsheng Wang
wangzs@nju.edu.cn

[1]College of Life Sciences, Nanjing University, Nanjing 210023, China
[2]Centre for Rainforest Studies, The School for Field Studies, Yungaburra, Qld 4884, Australia
[3]College of Life Sciences, Sichuan University, Chengdu 610065, China

Introduction

Together with Altingiaceae, Cercidiphyllaceae, and Daphniphyllaceae, Hamamelidaceae forms the woody clade within the order Saxifragales [1, 2]. As currently circumscribed, the extant family Hamamelidaceae comprises 28 genera and *c.* 120 species [3, 4]. Most of these genera are monotypic or oligo-species (e.g., *Chunia* Hung Ta Chang, *Embolanthera* Merrill, *Molinadendron* Peter Karl Endress, *Parrotia* Carl Friedrich von Meyer, *Shaniodendron* Mao Bin Deng, Hong Tu Wei & Xi Qu Wang and etc.), and many Hamamelidaceae species are categorized as endangered species by the IUCN Red List or China Species Red List [5, 6]. The Hamamelidaceae plants generally exhibit a pantropical intercontinental disjunct distribution pattern. Among these regions, Asia—particularly South China—serves as the modern distribution center and the diversification center of the family, as well as the center of its origin and evolutionary differentiation [7–9]. Fossil evidences indicated that Hamamelidaceae were widely distributed in the Northern Hemisphere in the Upper Cretaceous [10]. Most plants in the Hamamelidaceae family are trees or shrubs, typically inhabiting environments such as mountainous regions, stream valleys, and rocky slopes. Besides, Hamamelidaceae plants also present extensive morphological variation and remarkable diversity in their structural characteristics, for example, the inflorescence types (racemes, capitula, and spikes) and pollination types (entomophilous, ornithophilous, and anemophilous) [11, 12]. Additionally, the Hamamelidaceae species hold significant ecological and economic value. Species in this family are key components of forest subcanopy tree layers, often coexisting with other species or serving as locally dominant taxa. Members of the genera like *Chunia* Hung T. Chang, *Parrotiopsis* (Nied.) Camillo Karl Schneider, *Exbucklandia* (Robert Brown ex Griffith) Robert Wight & George Bentham, and *Mytilaria* Lecomte provide high-quality raw materials for furniture and construction. Plants from genera such as *Hamamelis* Gronovius ex Linné, *Corylopsis* Siebold & Zuccarini, *Mytilaria* L., and *Distylium* Siebold & Zuccarini have high medicinal value. *Hamamelis* Gronov. ex L. species are also used in cosmetic production; the tannins in *H. virginiana* L. hold the function of cleansing, anti-bacterial, anti-inflammatory, and skincare properties, making them common ingredients in personal care products. Moreover, most species in this family are highly valued for ornamental purposes because of their showy flowers, vibrant autumn foliage, and distinctive exfoliating bark. They are widely planted in landscape and urban greening projects across Eurasia [7–9, 13].

Since the initiation of molecular phylogenetic studies on Hamamelidaceae in the 1990s, considerable progress has been made in establishing the family's taxonomic framework [3, 4, 14–21]. However, the phylogenetic relationships among some species, genera, and tribes remain unresolved and contentious. First of all, the sister-group relationship between the monotypic genera *Shaniodendron* and *Parrotia* remains unresolved and debated [22–24]. Using nuclear ribosomal DNA (nrDNA) internal transcribed spacer (ITS) sequences, Li et al. [15] reconstructed a near-complete generic-level phylogeny of Hamamelidaceae (excluding *Chunia* and *Dicoryphe*), which provided moderate support (bootstrap, BS = 74) for the sister-group relationship between *Shaniodendron* and *Parrotia*. In contrast, based on the chloroplast *matK* gene, Li et al. [16] did not recover the sister-group relationship. The phylogenies inferred from the low-copy nuclear gene *GBSSI* and ITS provided strong support for the sister-group relationship between *Shaniodendron* and *Parrotia* (BS = 91) [3]. Meanwhile, the chloroplast gene trees did not support the relationship and instead recovered *Shaniodendron* as part of the monophyletic group comprising *Sinowilsonia*, *Distylium*, and *Parrotia* [3]. Using a combined dataset of six chloroplast markers (*atpB*, *matK*, *rbcL*, *atpB–rbcL*, *trnH–psbA*, and $trnL^{UAA}$–F^{GAA}) and nuclear ribosomal ITS sequences, Xiang et al. [4] reconstructed a comprehensive genus-level phylogeny of Hamamelidaceae and their analysis did not recover *Shaniodendron* and *Parrotia* as sister taxa. On the contrary, Bobrov et al. [18], utilizing five chloroplast regions (*matK*, *rbcL*, *trnL–trnF*, *psaA–ycf3*, and *psbA–trnH*) along with nrDNA ITS and 5.8S sequences, found high support for the sister-group relationship between *Shaniodendron* and *Parrotia* (BS = 87). Additionally, Zhang et al. [19] also fully supported the sister relationship of *Shaniodendron* and *Parrotia* (BS = 100) using 804 single-copy nuclear genes.

Secondly, the monophyly of the genus *Distyliopsis* Peter Karl Endress remains controversial. Based on nrDNA ITS sequences, Li et al. [25] failed to resolve the monophyly of the genus *Distyliopsis*, and suggested that further studies with expanded sampling were necessary. Research of Xiang et al. [4] showed that one representative species of *Distyliopsis*, *D. tutcheri*, formed a monophyletic clade with *Sycopsis*, while *Distyliopsis* itself did not constitute a monophyletic group. Bobrov et al. [18] included two *Distyliopsis* species—*D. tutcheri* and *D. dunnii*—but they did not cluster together; instead, both were nested within a monophyletic clade along with *Distylium* species. In contrast, Zhang et al. [19] strongly supported the monophyly of *Distylium* and *Distyliopsis* based on phylogenetic analyses of 804 single-copy nuclear genes. Thirdly, due to the scarcity of available material, the genus *Embolanthera* has been excluded from many molecular phylogenetic studies of Hamamelidaceae. To date, only the studies conducted by Xiang et al. [4] and Zhang et al. [19] have included this genus, and their results all supported that

Embolanthera clustered within a monophyletic group corresponding to the tribe Eustigmateae, rather than being placed in the tribe Loropetaleae suggested by Li [3].

Chloroplast is an important and dynamic energy converter organelle in plants and some algae, and it has independent genome system (chloroplast genome) with a circular double-stranded DNA molecule [26]. Besides, chloroplast genome is characterized by structural conservation, uniparental inheritance, low recombination rate, and evolves slowly in most angiosperms, which make it widely used in studies of molecular evolution and phylogenetics [26–28]. In recent years, with the rapid advancement of high-throughput sequencing technologies and significant reduction in sequencing costs, chloroplast phylogenomics based on genome skimming or deep genome skimming approaches has developed rapidly. Therefore, the resolution of many previously intractable phylogenetic relationships has been increased remarkably [29–32].

In this study, we adopted the classification system of Li [3] as the taxonomic framework of Hamamelidaceae and first sequenced and *de novo* assembled the complete chloroplast genomes of 37 representative species with comprehensive genus-level sampling within Hamamelidaceae. We subsequently investigated the mechanism of chloroplast genome evolution of Hamamelidaceae and developed genetic markers such as microsatellites and DNA barcodes for the family. Finally, we reconstructed the phylogenetic relationships of Hamamelidaceae based on the above 37 complete chloroplast genomes selected as ingroups and the 13 species from six genera of closely related families in the order Saxifragales used as outgroups. Our study will give a deep insight to the genome evolution and provide a robust phylogenetic framework for constructing a more resolved "Tree of Life" of Hamamelidaceae.

Results

Genome resequencing and chloroplast genome assembly results

Based on the Illumina HiSeq XTen platform, genome resequencing was performed on 49 newly-sequenced woody taxa within the order Saxifragales, generating a total of approximately 531.71 Gb of high-quality clean data (150 bp paired-end reads). For all samples (Supplementary Table 1), the percentage of bases with a Phred quality score ≥ 20 (Q20), corresponding to a sequencing error rate ≤ 1%. The GC content of the sequencing reads was also relatively high across all samples, indicating that the sequencing data were of excellent quality. For the two representative specimens, *Embolanthera spicata* and *Dicoryphe viticoides*, approximately 30 Gb of clean data were obtained for per individual, with Q20 values of 97.53% and 97.70%, and GC contents of 43.95% and 43.48%, respectively.

Chloroplast genome *de novo* assembly for all 49 newly-sequenced woody Saxifragales taxa was conducted using a combination of GetOrganelle v.1.4.3 and CLC Genomics Workbench v.8.1. As a result, complete circular chloroplast genome sequences were successfully reconstructed for all individuals and subsequently used for downstream analyses. The above 49 newly-sequenced complete chloroplast genomes of woody Saxifragales (including 37 representative species of Hamamelidaceae) generated in this study were deposited in the GenBank with accession numbers ON693853, ON729341 to ON729387 (Supplementary Table 2).

Chloroplast genome characteristics of Hamamelidaceae

The complete chloroplast genomes of 37 species representing all 28 genera of Hamamelidaceae exhibited a typical quadripartite circular structure (Fig. 1). The genome sizes ranged from 158,090 bp (*Disanthus cercidifolius* var. *longipes*) to 160,814 bp (*Exbucklandia tonkinensis*) (Supplementary Table 2). The length of the large single-copy (LSC) region varied between 87,057 bp (*Disanthus cercidifolius* var. *longipes*) and 89,413 bp (*Mytilaria laosensis*), while the small single-copy (SSC) region ranged from 18,094 bp (*Rhodoleia championii*) to 18,911 bp (*Exbucklandia populnea*). The length of the inverted repeat (IR) regions spaned from 26,158 bp (*Parrotiopsis jacquemontiana*) to 26,429 bp (both *Exbucklandia tonkinensis* and *Neostrearia fleckeri*). The overall GC content across these plastomes ranged from 37.7% (*Rhodoleia championii*) to 38.2% (*Matudaea trinervia* and *Molinadendron sinaloense*) (Supplementary Table 2).

The chloroplast genomes of the 37 species representing all 28 genera of Hamamelidaceae exhibited highly conserved gene content and order. Based on the annotation results, these genomes can be categorized into three types (Table 1 and Supplementary Table 2): 1) Type I: the plastomes of 23 species from 18 genera contained a total of 133 genes, including 115 unique genes and 18 duplicated genes. The 115 unique genes comprise 80 protein-coding genes (PCGs), 30 tRNA genes, 4 rRNA genes, and one pseudogene ($^{\Psi}$*ycf1*). The pseudogene $^{\Psi}$*ycf1* arises due to incomplete duplication of the *ycf1* gene, which spaned the SSC/IRa boundary, resulting in a truncated copy in the corresponding IRb region. 2) Type II: the plastomes of 14 species from 11 genera contained 134 genes, consisting of 116 unique genes and 18 duplicated genes. The 116 unique genes include 80 PCGs, 30 tRNA genes, 4 rRNA genes, and two pseudogenes ($^{\Psi}$*ycf1* and $^{\Psi}$*rps19*). The formation of $^{\Psi}$*rps19* is similarly attributed to incomplete duplication of the *rps19* gene, which spaned the LSC/IRb junction, leading to a partial, non-functional copy in the IRa region. 3) Type III: the plastome of *Chunia* Hung T.

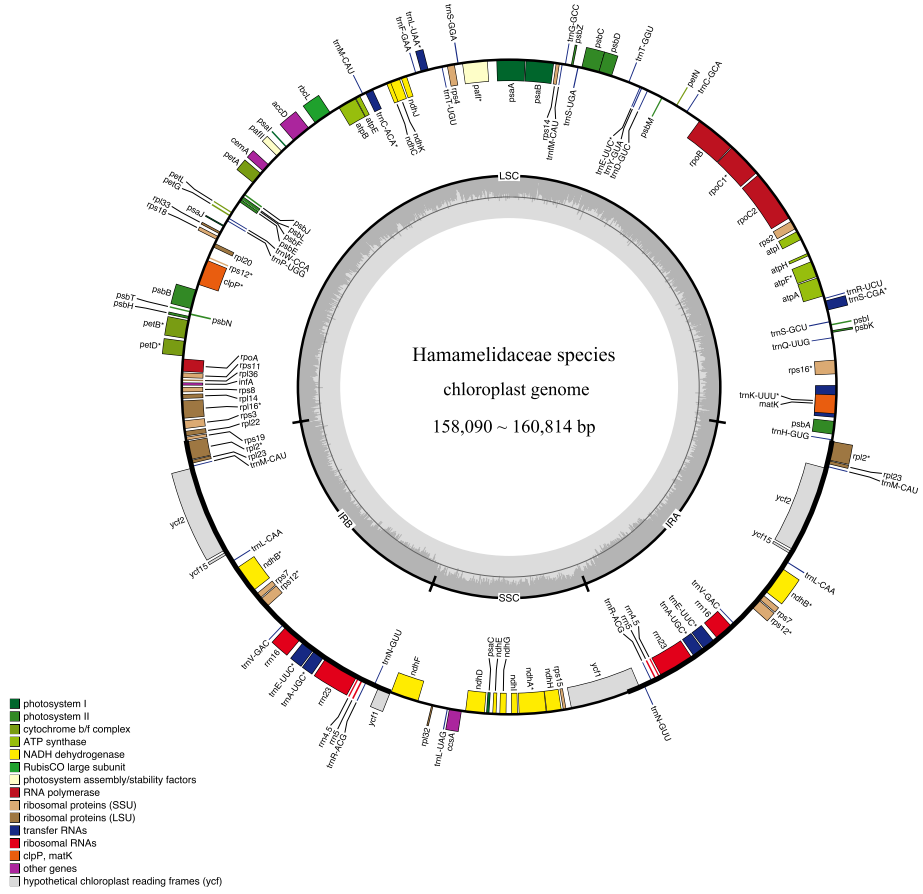

Fig. 1 Gene map of chloroplast genomes of 37 species of Hamamelidaceae. Genes shown on the outside of the circle are transcribed clockwise, and genes inside are transcribed counter-clockwise. Genes belonging to different functional groups are color-coded. The darker gray in the inner corresponds to GC content, and the lighter gray corresponds to AT content

Chang also contained 133 genes, with 115 unique genes and 18 duplicated genes. However, its 115 unique genes included 78 PCGs, 30 tRNA genes, 4 rRNA genes, and three pseudogenes (one Ψ*ycf1* and two Ψ*ycf15*). The *ycf15* gene is pseudogenized due to the presence of premature stop codons within its sequence, leading to two Ψ*ycf15* copies. In summary, variation in chloroplast genome gene composition within Hamamelidaceae was mainly attributable to differences in the number and type of pseudogenes.

In addition, a total of 18 intron-containing genes were identified in the chloroplast genomes of Hamamelidaceae species. Among them, 15 genes contained a single intron, including nine protein-coding genes (*atpF, ndhA, ndhB, petB, petD, rpl2, rpl16, rpoC1,* and *rps16*) and six tRNA genes (*trnG-GCC, trnK-UUU, trnL-UAA, trnV-UAC, trnI-GAU,* and *trnA-UGC*). Only three protein-coding genes (*ycf3, clpP,* and *rps12*) were found to contain two introns. Notably, *rps12* is a trans-spliced gene, with its 5' exon located in the LSC region and its 3' exons and introns located within the IR regions (Table 1).

Comparative chloroplast genome analysis of Hamamelidaceae

Global chloroplast genome sequence comparisons within 37 congeneric Hamamelidaceae species were conducted using the mVISTA software under the Shuffle-LAGAN mode and with *Exbucklandia tonkinensis* as the reference. The visualization results indicated that Hamamelidaceae chloroplast genomes were conserved with a high degree of similarity and synteny at the genome-scale level. However, we also found the sequence mutations between Exbucklandioideae and the other sub-families were relatively larger. Subfamilies Disanthoideae and Hamamelidoideae, and the tribe Hamamelideae and Fothergilleae had a higher degree of similarity in genome sequences, indicating the close phylogenetic relationships among those taxa. Overall, the IR regions were more conserved than the LSC and SSC regions, and the sequence variation in the non-coding regions was higher than that of genes in coding regions (Supplementary Fig. 1). Synteny analysis using the Mauve software revealed relatively conserved chloroplast genome structures within Hamamelidaceae, while few re-arrangements in the LSC

Table 1 Gene list of chloroplast genomes of 37 Hamamelidaceae woody taxa

Groups of genes	Names of genes
Ribosomal RNAs	*rrn16* (×2), *rrn23* (×2), *rrn4.5* (×2), *rrn5* (×2)
Transfer RNAs	*trnK-UUU*[a], *trnQ-UUG*, *trnS-GCU*, *trnG-GCC*[a], *trnR-UCU*, *trnC-GCA*, *trnD-GUC*, *trnY-GUA*, *trnE-UUC*, *trnT-GGU*, *trnS-UGA*, *trnG-UCC*, *trnfM-CAU*, *trnS-GGA*, *trnT-UGU*, *trnL-UAA*[a], *trnF-GAA*, *trnV-UAC*[a], *trnM-CAU*, *trnW-CCA*, *trnP-UGG*, *trnH-GUG* (×2), *trnI-CAU* (×2), *trnL-CAA* (×2), *trnV-GAC* (×2), *trnI-GAU*[a] (×2), *trnA-UGC*[a] (×2), *trnR-ACG* (×2), *trnN-GUU* (×2), *trnL-UAG*
Photosystem I	*psaB*, *psaA*, *psaI*, *psaJ*, *psaC*
Photosystem II	*psbA*, *psbK*, *psbI*, *psbM*, *psbD*, *psbC*, *psbZ*, *psbJ*, *psbL*, *psbF*, *psbE*, *psbB*, *psbT*, *psbN*, *psbH*
Cytochrome	*petN*, *petA*, *petL*, *petG*, *petB*[a], *petD*[a]
ATP synthase	*atpA*, *atpF*[a], *atpH*, *atpI*, *atpE*, *atpB*,
Rubisco	*rbcL*
NADH dehydrogenase	*ndhJ*, *ndhK*, *ndhC*, *ndhB*[a] (×2), *ndhF*, *ndhD*, *ndhE*, *ndhG*, *ndhI*, *ndhA*[a], *ndhH*
ATP-dependent protease subunit P	*clpP*[b]
Chloroplast envelope membrane protein	*cemA*
large units	*rpl33*, *rpl20*, *rpl36*, *rpl14*, *rpl16*[a], *rpl22*, *rpl2*[a] (×2), *rpl23* (×2), *rpl32*
small units	*rps16*[a], *rps2*, *rps14*, *rps4*, *rps18*, *rps12*[b] (×2), *rps11*, *rps8*, *rps3*, *rps19*, *rps7* (×2), *rps15*
RNA polymerase	*rpoC2*, *rpoC1*[a], *rpoB*, *rpoA*
Miscellaneous proteins	*matK*, *accD*, *ccsA*
Hypothetical proteins & conserved reading frames	*ycf3*[b], *ycf4*, *ycf2* (×2), *ycf1*
Pseudogenes	[Ψ]*ycf15* (×2), [Ψ]*ycf68* (×2), [Ψ]*infA*

[a]indicates the genes containing a single intron

[b]indicates the genes containing two introns

(×2) indicates genes duplicated in the IR regions; pseudogene is represented by Ψ

regions of the chloroplast genomes across 37 Hamamelidaceae species (Supplementary Fig. 2).

Comparative analysis of IR boundaries indicated that border genes were identical among 37 Hamamelidaceae chloroplast genomes, but slight differences existed in length of these genes (*rps19*, *ndhF* and *ycf1*) and relative positions of these genes to the boundaries (Fig. 2). Specifically, in the junction of the LSC/IRb (*JLB*), the genes *rps19* and *rpl2* were located. Gene *rps19* with 279 bp in length exhibited two patterns among the 37 Hamamelidaceae species: (1) in 15 species, *rps19* extended into IRb by 2 bp, with exception of *Dicoryphe viticoides* (29 bp) and *Maingaya malayana* (66 bp); (2) in 22 species, *rps19* was entirely in the LSC region with 1–19 bp away from the *JLB*. Gene *rpl2* was placed completely in the IRb region. Regarding the IRb/SSC junction (*JSB*), the truncated *ycf1*

Fig. 2 Comparison of the boundaries of the large single-copy (LSC), the small single-copy (SSC) and inverted repeat (IR) regions among the 37 Hamamelidaceae chloroplast genomes. *JLB*: junction line between LSC and IRb; *JSB*: junction line between SSC and IRb; *JSA*: junction line between SSC and IRa; *JLA*: junction line between LSC and IRa. The figure does not scale with the actual sequence length and only shows relative changes at or near the IR/SC boundaries

(pseudogene: $^{\Psi}$*ycf1*) and *ndhF* were adjacent. The majority of $^{\Psi}$*ycf1* pseudogenes (985–1,155 bp in length) was located in the IRb region, with 1 bp apart from the *JSB* boundary. The *ndhF* gene (2,231–2,240 bp in length) displayed two configurations: one was the joint location in the SSC region; the other was the existence of overlap between *ndhF* and $^{\Psi}$*ycf1*, with which *ndhF* expanded into the IRb region (0–22 bp). In the *JSA* boundary, the gene *ycf1* crossed over the *JSA* with a length of 4,439–4,594 bp and 986–1,156 bp in the SSC and IR region, respectively. The IRa/LSC junction (*JLA*) was located between the other *rpl2* and *trnH-GUG*. This *rpl2* gene completely located in the IRa region and the gene *trnH-GUG* with 74 bp in length was 7–228 bp away from the *JLA* boundary. Besides, some truncated *rps19* (pseudogene: $^{\Psi}$*rps19*) exhibited in the *JLA* boundary in some species of Hamamelidaceae (Fig. 2).

Enrichment of chloroplast genetic resources of Hamamelidaceae

A total of 6,084 chloroplast simple sequence repeats (cpSSRs) including 5,293 mononucleotide (A, C, T, G), 321 dinucleotide (AT, AG, TA, TC, CT, CG), 160 trinucleotide (GAA, TTA, ATA, AAT, ATT, TCA, TAA, TAT, CAG, TTC), 247 tetranucleotide (ATAC, TGAA, TTCT, TTTC, AAAT, ATAA, TTTA, AGAA, ATCT, TAAA, TTAA, GAAA, AATA, CAAA, TTGA, ATTT, TTGT, GATA, TATT, AAAG, TTAT, AGAT, CAAT, TCTA, ATAG, TAAT), 56 pentanucleotide (TATTT, TTCTA, TCATA, TTCTA, TTCTT, TATAA, AAATG, ATTTT, ATATA, TAGAA), and 7 hexanucleotide (CTTTTT, CTAAGA, TCTCGG, ATCTTA, GAGAAG, TCCTTC) repeats were identified within the 37 Hamamelidaceae species (Fig. 3A, Supplementary Table 3). The majority of cpSSRs were located in the LSC regions (76.6%, Fig. 3B, Supplementary Table 3), with mononucleotide cpSSRs (predominantly A/T repeats) comprising 87.0% of the total. The number of cpSSRs per species ranged from 136 (*Disanthus cercidifolius*) to 179 (*Dicoryphe viticoides*) (Supplementary Table 3). Notably, most cpSSR loci were located in the IGS regions (81.7%), followed by CDS (5.6%) and introns (12.8%) (Fig. 3C). Analysis of the scattered sequence repeats via REPuter identified 1,443 scattered repeats totally across the 37 Hamamelidaceae species, ranged from 28 long repetitive sequences in *Ostrearia australiana* to 144 in *Dicoryphe viticoides* (Supplementary Table 4). Specifically, 661 forward repeat sequences (49.3%), 29 reverse repeat sequences (2.2%), 3 complementary repeat sequences (0.2%), and 648 palindromic repeat sequences (48.3%) (Fig. 3D).

Nucleotide variability (π) within the 37 Hamamelidaceae species was analyzed by sliding window analysis using multiple-sequence alignment. A total of 195 gene fragments were detected in the chloroplast genomes of 37 Hamamelidaceae species, with a mean nucleotide diversity value of 0.01646. Four nucleotide polymorphism hotspots (DNA barcodes) were discovered in four genes (*ndhC/rps18/trnT-GGU/trnR-UCU*) with π values above 0.045, which were located in single-copy regions (Fig. 4 and Supplementary Table 5).

Evolutionary mechanism and pan-genomic analysis of Hamamelidaceae

We calculated the relative synonymous codon usage (RSCU) value of 37 Hamamelidaceae chloroplast genomes. The results revealed that the Hamamelidaceae chloroplast genomes shared 59 codons, among which 29 codons had RSCU values greater than 1, predominantly ending with adenine (A) and uracil (U). Specifically, 20 codons (UUA, AUU, GUU, GUA, UCU, CCU, ACU, GCU, CAU, CAA, AAU, AAA, GAU, GAA, UGU, CGU, CGA, AGA, GGU, GGA) exhibited RSCU values > 1.3, five codons (UUU, UUG, CUU, UCA, ACA) had RSCU values between 1.2 and 1.3, and three codons (AGU, CCA, GCA) displayed RSCU values ranging from 1.0 to 1.2 (Supplementary Fig. 3).

The analysis of selection pressure on 76 protein-coding genes in the plastomes of Hamamelidaceae using CODEML revealed that the *clpP* gene was under positive selection (adaptive selection, dN/dS > 1) (Supplementary Fig. 4, Supplementary Table 6). Apart from this gene, the others may be subject to neutral or purifying selection.

Based on the 0/1 matrix of chloroplast genes from 37 species (totaling 137 gene loci), the pan-core accumulation analysis revealed rapid saturation of the chloroplast gene pool. Specifically, the pan gene number converges rapidly to 137 with increasing sample size, while the core gene number converges to 131. The strict core gene proportion is 95.6% (131/137), with only five genes—*rps19*_copy2, *trnE-UUC*_copy2, *trnE-UUC*_copy3, *trnI*, and *trnI*_copy2—being accessory (present in 2–36 species) and *trnH-GUG*_copy2 being unique (present only in the species *Maingaya malayana*) (Supplementary Fig. 5). These findings suggested that the chloroplast gene content of this family is highly conserved overall, with the gene pool approaching a "closed" state.

Phylogenetic relationships of Hamamelidaceae

The complete chloroplast genome sequences were aligned using MAFFT v.7 [33], and ambiguously aligned regions were trimmed. The final concatenated alignment matrix comprised 176,275 bp. Substitution saturation was assessed using DAMBE v.5.3.74 [34], and the results indicated that the dataset is not saturated ($Iss < Iss.c$, $P = 0.0000 < 0.001$), supporting its suitability for subsequent phylogenetic analysis. In addition, 76 concatenated protein-coding genes (PCGs) were aligned with MAFFT v.7 [33], and ambiguous sites were removed, resulting in

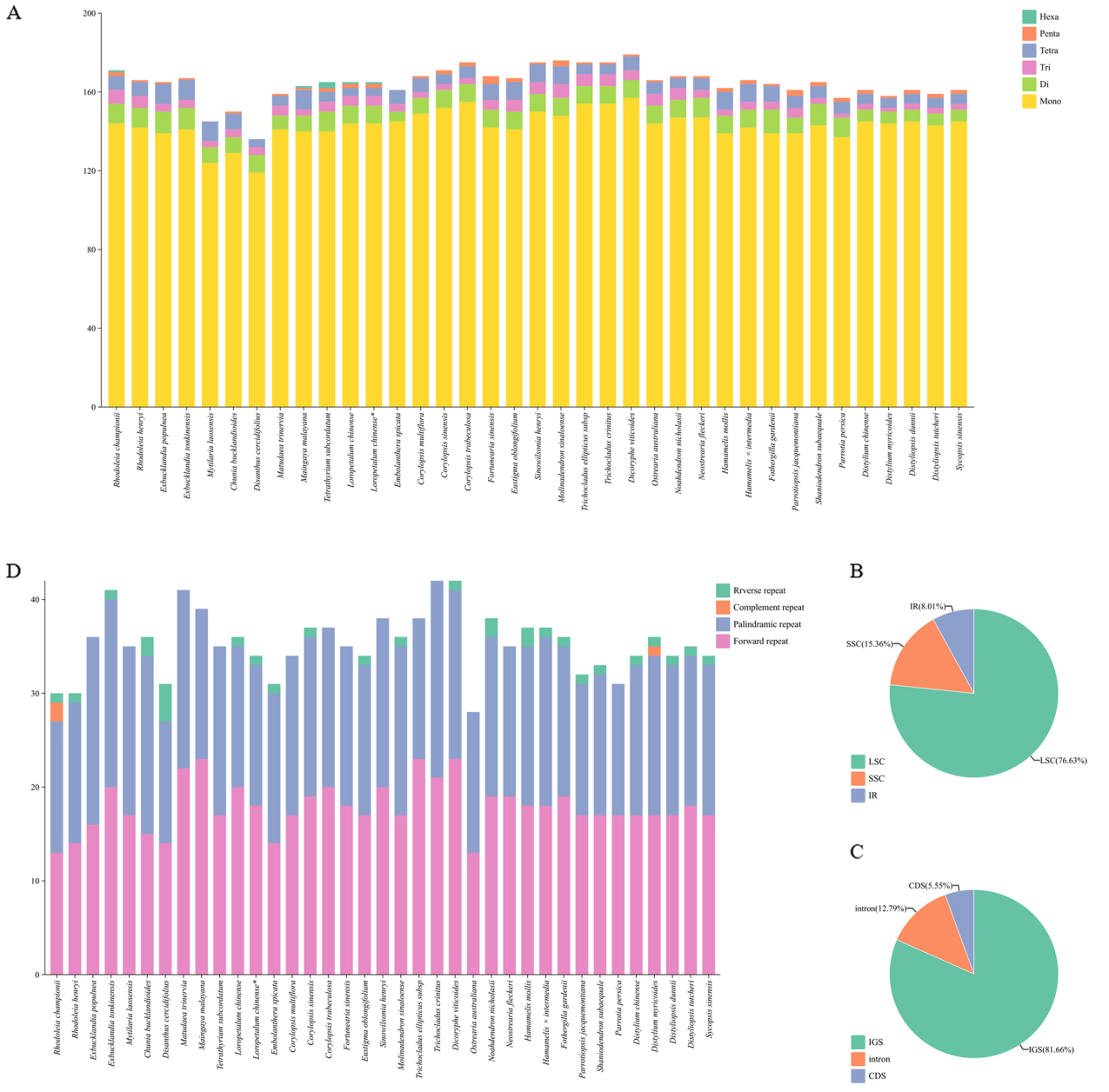

Fig. 3 Comparative analysis of chloroplast genome repeated sequences of 37 species from Hamamelidaceae. **A**: The numbers of the six cpSSR types; **B**: The numbers of cpSSRs distributed in different copy regions; **C**: The numbers of cpSSRs distributed in different gene regions; **D**: The numbers of the four long repeat sequence types

an alignment matrix of 69,090 bp. Substitution saturation testing with DAMBE v.5.3.74 [34] similarly showed that the sequences were not saturated (Iss < Iss.c, $P = 0.0000 < 0.001$), indicating that this dataset is also appropriate for phylogenetic reconstruction.

Phylogenetic trees were reconstructed using both Maximum Likelihood (ML) and Bayesian Inference (BI) methods based on the two datasets (complete chloroplast genomes and 76 concatenated PCGs). The resulting four phylogenetic trees shared identical topologies, with high support values at nearly all nodes (Fig. 5 and Supplementary Fig. 6). For the complete plastome dataset, the median bootstrap support (BS) and posterior probability (PP) values were 100/1, the average values were 98/1, and the standard deviations were 5.74 for BS and 0.002 for PP. For the PCG dataset, the median BS/PP values were also 100/1, with average values of 99/1 and standard deviations of 2.66 for BS and 0.003 for PP. These results indicated that both datasets provided highly consistent and strongly supported phylogenetic signals, validating their reliability for resolving evolutionary relationships within Hamamelidaceae.

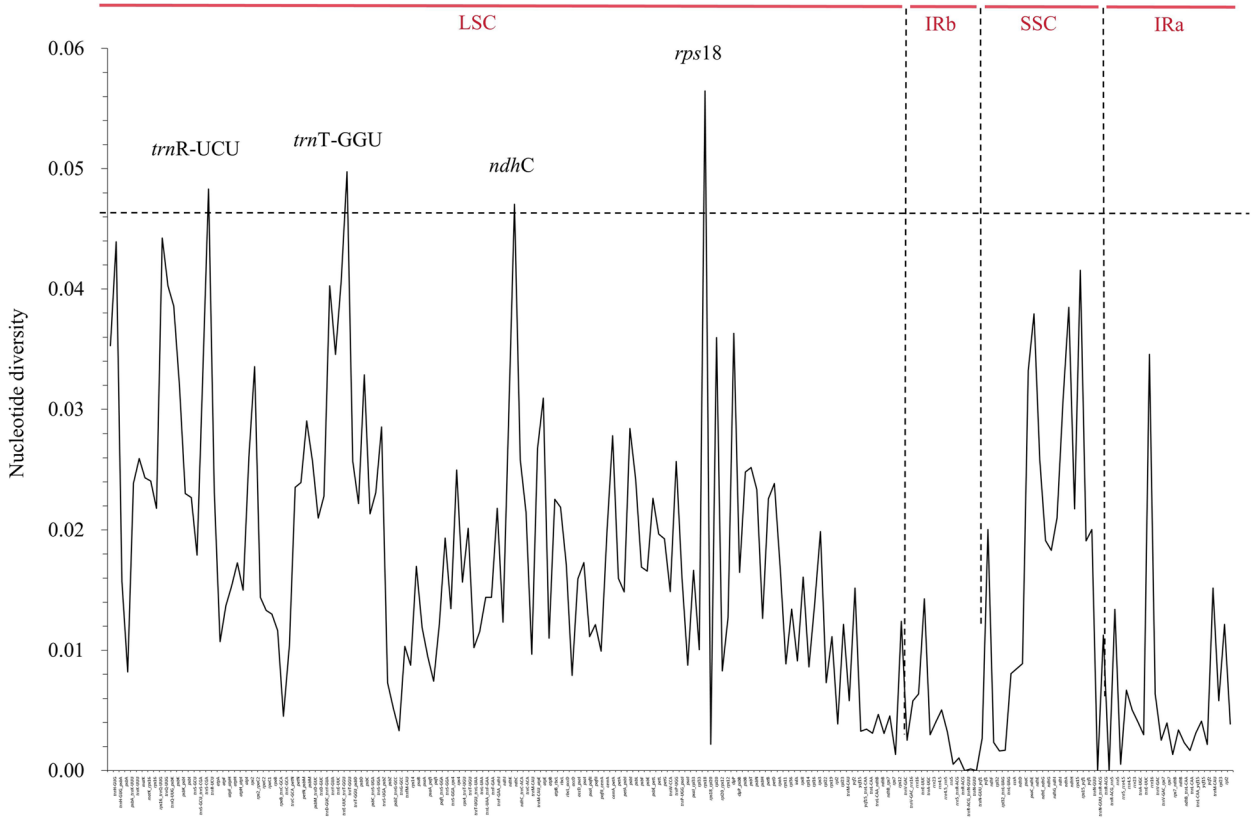

Fig. 4 Comparison of nucleotide variability (π) values among the 37 Hamamelidaceae chloroplast genomes (window length: 600 bp; step size: 200 bp)

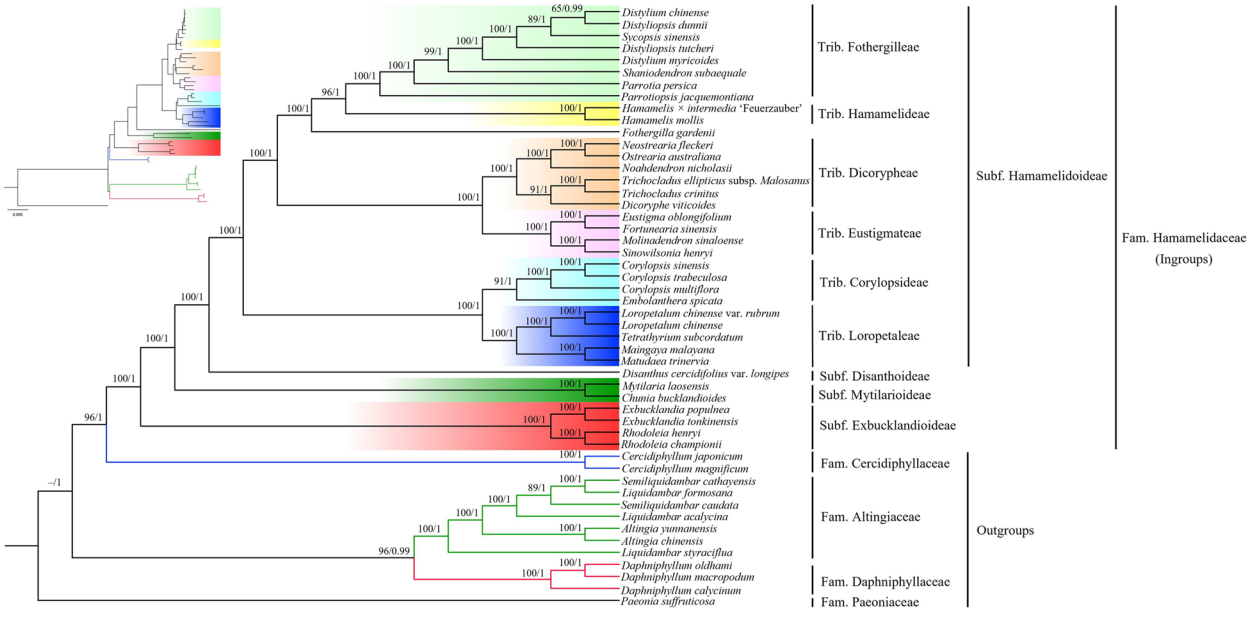

Fig. 5 Phylogeny of Hamamelidaceae based on the whole chloroplast genome via ML and BI methods. The numbers at each node represent ML bootstrap support (BS) and BI posterior probability (PP), ML and BI tree show totally congruent topologies

Regarding the phylogenetic relationships among woody lineages within Saxifragales, the above-mentioned topologies strongly supported Hamamelidaceae and Altingiaceae as sister groups (whole plastome tree:BS/PP = 96/1; PCGs tree: BS/PP = 100/1). Daphniphyllaceae and Cercidiphyllaceae formed a monophyletic clade (whole plastome tree BS/PP = 96/0.99; PCGs tree BS/PP = 100/1), which showed a closer relationship with

Paeoniaceae. The majority of extant genera of Hamamelidaceae formed a well-supported monophyletic clade, and the intrafamilial classification was largely congruent with the system proposed by Li [3]. All major clades received strong support (whole plastome tree: BS/PP = 100/1; PCGs tree: BS/PP = 100/1). The four subfamilies of Hamamelidaceae were each recovered as monophyletic clades. Exbucklandioideae was the earliest-diverging lineage, followed sequentially by Mytilarioideae, Disanthoideae, and Hamamelidoideae.

Within Hamamelidoideae, the phylogenetic relationships among tribes and genera showed several incongruences compared with the classification system of Li [3], as detailed below: 1) *Embolanthera* Merr. formed a monophyletic clade with members of tribe Corylopsideae (whole plastome tree: BS/PP = 91/1; PCGs tree: BS/PP = 98/1), rendering tribe Loropetaleae, as circumscribed by Li [3], paraphyletic; 2) Among the five genera endemic to the Southern Hemisphere, *Trichocladus* Pers. and *Dicoryphe* Thouars formed one clade (BS/PP = 91/1 for whole plastome tree; 97/1 for PCGs tree), while *Ostrearia* Baill., *Noahdendron* P.K. Endress, B. Hyland & R. Tracey, and *Neostrearia* L.S. Sm. formed another well-supported clade (BS/PP = 100/1 for both datasets). Within this latter group, *Neostrearia* and *Ostrearia* constituted a monophyletic subclade (BS/PP = 100/1 for both datasets). These two major clades were themselves sister groups (BS/PP = 100/1 for both trees); 3) *Fothergilla* L. was not placed within tribe Fothergilleae, contradicting its position in Li's [3] classification (BS/PP = 100/1 for both datasets); 4) *Shaniodendron* M.B. Deng, H.T. Wei & X.Q. Wang and *Parrotia* C.A. Mey. did not form a sister relationship (BS/PP = 100/1 for both datasets); instead, *Shaniodendron* groups within a monophyletic clade alongside *Distylium*, *Parrotiopsis*, and *Distyliopsis* (BS/PP = 99/1 for both datasets); 5) Representatives of *Distylium* Siebold & Zucc. and *Distyliopsis* P. K. Endress did not each form monophyletic groups.

Discussion

Chloroplast genome evolution of Hamamelidaceae

Chloroplast genome is crucial for disentangling and advancing our understanding of the plant evolutionary mechanisms and adaptation [26–28]. In this study, we investigated and charactered the chloroplast genomes of 37 Hamamelidaceae species at the whole genus level to capture the intergenus/intragenus variation and evolution of Hamamelidaceae chloroplast genomes. Our results showed a generally conserved genome structure, gene arrangement, gene and GC content, codon usage bias, and core genes in Hamamelidaceae. Nevertheless, differences in pseudogenes and positive selections in the chloroplast genome evolution of Hamamelidaceae were identified. We will discuss them below:

1. Pseudogenes evolution in Hamamelidaceae.

Pseudogenes possess biological functions and serve as raw material for the emergence of new genes and functions, playing important roles in gene expression regulation and species evolution [35–38]. The identification of three pseudogenes ($^{\Psi}$*ycf1*, $^{\Psi}$*rps19* and $^{\Psi}$*ycf15*) in Hamamelidaceae's cp genomes provides a unique perspective on the genomic evolutionary dynamics of this family. Specifically, pseudogene $^{\Psi}$*ycf1* was found in 16 genera of Hamamelidacea, $^{\Psi}$*rps19* was detected in 11 genera, while $^{\Psi}$*ycf15* only spotted in the genus of *Chunia*, indicating a genus specificity of pseudogenes among Hamamelidacea's species.

In this study, $^{\Psi}$*ycf1* and $^{\Psi}$*rps19* were identified as pseudogenes due to the presence of a premature stop codon, while $^{\Psi}$*ycf15* contained multiple internal stop codons. The *ycf1* gene is crucial for plant growth and reproduction as it encodes the Tic214 protein in the chloroplast genome [39, 40]. The *rps19* is involved in the self-replication of chloroplasts and vital for ribosome biogenesis, impacting proper gene expression and cell growth [41]. The *ycf15* gene belongs to a group of chloroplast genes with unknown function, yet due to its distinct sequence structure and abundance of phylogenetically informative sites, *ycf15* holds considerable value in plant phylogenetics and species identification [42–44]. Recent review pointed that pseudogenes could be functionally reactivated under specific conditions to exert regulatory roles in shaping species-specific adaptations [45]. We therefore inferred that pseudogenes of Hamamelidacea may possess certain functionalities and could acquire new functions through evolutionary processes. However, more study should be conducted to demonstrate the above hypothesis.

2. Positive selection of *clpP* in Hamamelidaceae.

In this investigation, our analysis revealed that among the protein-coding genes of Hamamelidaceae, *clpP* gene was found to be under positive selection ($dN/dS > 1$), indicating ongoing adaptive evolutionary processes within Hamamelidaceae.

The *clpP* gene encodes a chloroplast protease subunit that regulates protein homeostasis and photosystem repair, which may be related to the plant's ability to adapt to environmental stresses such as drought and salinity [46]. Given the heterogeneous habitats of Hamamelidaceae, the *clpP* gene's positive selection-driven variation optimizes photosynthetic efficiency under adverse conditions, providing a better understanding of the molecular mechanisms of plant adaptive evolution of this family.

Chloroplast genome-based phylogeny of Hamamelidaceae

Accurate phylogenetic relationships among clades provide a fundamental basis for understanding species evolution and the formation of floristic, and they are essential for conducting interdisciplinary and integrative research [47, 48]. As a representative component of the East Asian flora, Hamamelidaceae possesses significant ecological and economic value. Its origin and evolutionary history have long been of great interest to evolutionary biologists.

In this study, we reconstructed—for the first time—the phylogenetic relationships among all genera of Hamamelidaceae and other woody families within Saxifragales, based on the complete chloroplast genome sequences. With regard to the intrageneric phylogenetic relationships within Hamamelidaceae, our results supported the monophyly of the majority of extant genera in the family. The four subfamilies, as well as the major tribes within Hamamelidoideae, are largely consistent with the classification framework proposed by Li [3]. However, certain phylogenetic relationships inferred in this study conflicted with previous findings, and these inconsistencies were discussed in detail below:

1. Evidence from this study indicated that tribe Loropetaleae, as circumscribed by Li [3], was not monophyletic. Specifically, *Embolanthera* Merr., originally assigned to Loropetaleae, was strongly supported as forming a monophyletic clade with members of tribe Corylopsideae (whole plastome tree: BS/PP = 91/1; PCGs tree: BS/PP = 98/1). Our findings were consistent with those of Xiang et al. [4] and Zhang et al. [19]. Morphological evidence further supported our results: all other members of Loropetaleae (except *Embolanthera*) possess disc-shaped lobes that produce nectar [49], a trait not observed in *Embolanthera*. Moreover, species of *Corylopsis* and *Loropetalum* (both in Corylopsideae) share several morphological features, including being deciduous shrubs or small trees, having alternate, ovate to obovate leaves, spathulate petals, five stamens with two-loculed anthers, and a single pendulous ovule per locule. Based on both molecular and morphological evidences, we suggested that the tribal placement of *Embolanthera* should be re-evaluated.
2. The phylogenetic reconstruction of the five genera endemic to the Southern Hemisphere in this study differed from the results of previous studies [3, 4, 15, 18], but was consistent with Zhang et al. [19] based on 804 single-copy nuclear genes, all of which strongly supported a sister-group relationship between *Neostrearia* and *Ostrearia* (whole plastome tree: BS/PP = 100/1; PCGs tree: BS/PP = 100/1). Given that the sister relationship between *Neostrearia* and *Noahdendron* had received low support with limited phylogenetic signals in most earlier studies, and in light of morphological evidence—particularly the similarity in pollen morphology between *Neostrearia* and *Ostrearia* [50]—this study likewise concluded that the phylogenetic relationships among these three genera remain unresolved.
3. Our results showed that the representative species of *Fothergilla*, *Fothergilla gardenii*, did not fall within tribe Fothergilleae as defined in the classification system of Li [3]. However, this finding revealed a clear case of cytonuclear discordance when compared with the results based on nuclear genomic data [19]. The tetraploid hybrid origin of *Fothergilla gardenii* [51] may explain this discrepancy. According to the research of Qi et al., a chloroplast capture event during ancient hybridization could have led to the anomalous placement of this genus in the chloroplast-based phylogenetic tree [51]. However, more direct and robust evidences should be provided to accurately explain the chloroplast topology in the future study.
4. The sister-group relationship between *Shaniodendron* and *Parrotia* was not supported by the chloroplast genome phylogeny in this study. Instead, *Shaniodendron* was strongly supported as forming a monophyletic clade with *Distylium*, *Parrotiopsis*, and *Distyliopsis*. This result was consistent with previous studies based on chloroplast DNA fragments [3, 16]. Li [3] also proposed a possible inference for this pattern, assuming that *Shaniodendron* may have undergone ancient hybridization, leading to a chloroplast capture event from the ancestral lineage of the *Distylium*–*Parrotiopsis*–*Distyliopsis* clade. In contrast, nuclear phylogenomics fully supported the sister relationship of *Shaniodendron* and *Parrotia* (BS = 100) using 804 single-copy nuclear genes [19] exhibiting a pronounced cytonuclear discordance. Therefore, more robust evidences should be collated to resolve their sister-group relationship in the future study.
5. To explain the non-monophyly observed for *Distylium* and *Distyliopsis* in our plastome phylogenetic analyses, we proposed that these two genera were closely related, as indicated by the short evolutionary branch lengths between them in the chloroplast genome phylogeny. Moreover, most species of both genera exhibit sympatric distributions, which increases the likelihood of hybridization and introgression events. Meanwhile, the phylogenetic analysis based on 804 single-copy nuclear genes strongly supported the monophyly of *Distylium* and *Distyliopsis* [19]. This result was

incongruent with the topology inferred from our chloroplast genome data and also represented a typical case of cytonuclear discordance. We inferred that the likely cause of this discordance was the close phylogenetic relationship between the two genera, which may have facilitated chloroplast capture events through hybridization or introgression. However, more direct and roboust evidences should be provided to correctly understand the chloroplast topology in the future study.

Overall, in this study, while plastome-scale data substantially improve phylogenetic resolution compared to single- or few-locus analyses, they may still be insufficient to completely resolve backbone relationships in cases of rapid radiations or ancient divergences. We therefore proposed that integrating nuclear genomic data and morphological evidence would likely provide greater power to disentangle the deeper phylogenetic and evolutionary relationships within Hamamelidaceae in the future research.

Materials and methods

Hamamelidaceae taxon sampling, DNA extraction and genome resequencing

A comprehensive taxonomic sampling of 50 accessions was conducted for the phylogenomic study of Hamamelidaceae, including 37 species representing all 28 currently recognized genera within the family [2]. These species were selected as ingroups. Additionally, 13 species from four closely related families in the order Saxifragales served as outgroups. Specifically, 12 newly-sequenced woody taxa belonged to the families of Altingiaceae, Cercidiphyllaceae, and Daphniphyllaceae. Besides, the one chloroplast genome data of *Paeonia suffruticosa* (Paeoniaceae) was downloaded from NCBI (NC_037879) [1, 2]. The voucher specimens were identified by David Yue Ping Ting and Pan Li, and had been deposited at the Herbarium of Nanjing University (HNU), with deposition numbers provided in Supplementary Table 1.

High-quality genomic DNA of 49 newly-sequenced woody taxa within the order Saxifragales was extracted from the silica-gel-dried leaves using a modified CTAB-based method [52]. The extracted DNA was then used to construct paired-end sequencing libraries with an insert size of 500–600 bp, following the Illumina library preparation protocol. Genome resequencing (150 bp paired-end reads) was performed on the Illumina HiSeq XTen platform at the Beijing Genomics Institute (BGI, Shenzhen, China). The raw sequencing data underwent trimming and filtering to remove adapters, low-quality reads, undersized inserts, and duplicate reads. This was done using TRIMMOMATIC v.0.39 [53], with a sliding window size of 4 and a minimum phred score of 20.

Chloroplast genome *de novo* assembly and annotation

Chloroplast genomes of the above 49 newly-sequenced species were *de novo* assembled from clean sequencing data using GetOrganelle v.1.7.6.1 [54], with *Shaniodendron subaequale* plastome (GenBank ID: NC_037243.1) as the reference genome. Before *de novo* assembly, we filtered and removed the adapter sequences, low-quality reads ($Q < 20$, 0.01 probability error), low coverage regions or ambiguous bases using NGS QC Tool Kit v.2.3.3 [55]. The core command used in GetOrganelle v.1.7.6.1 was: get_organelle_from_reads.py -s NC_037243.1 fasta –1 Species_1.fq.gz –2 Species_2.fq.gz -o Species_plastome_output -R 15 -k 21,45,65,85,105 -F embplant_pt. Additionally, the chloroplast genomes of *Embolanthera spicata* Merr. and *Dicoryphe viticoides* Baker were further *de novo* assembled using CLC Genomics Workbench v.8.1 (http://www.clcbio.com; CLC Bio, Denmark).

The assembled chloroplast genome sequences of 49 newly-sequenced species were then annotated using the Plastid Genome Annotator (PGA) pipeline [56]. The presence and structure of transfer RNA (tRNA) genes were further verified using tRNAscan-SE [57]. The annotated genomes were then imported into Geneious Prime (http://www.geneious.com/; Biomatters Ltd., Auckland, New Zealand), where manual adjustments were made based on *Shaniodendron subaequale* plastome (GenBank ID: NC_037243.1) to refine the positions of start and stop codons, as well as exon–intron boundaries of each gene.

To visualize the circular genome maps, the annotated sequences were submitted to Organellar Genome DRAW v1.3.1 [58] (http://ogdraw.mpimp-golm.mpg.de/). Finally, the annotated plastome sequences were submitted to the NCBI GenBank database (Supplementary Table 1).

Comparative chloroplast genome analysis of Hamamelidaceae

To clarify the level of sequence divergence of Hamamelidaceae, we conducted the chloroplast genomes comparison across 37 Hamamelidaceae species using mVISTA program (http://genome.lbl.gov/vista/mvista/, accessed on 18 November 2023) [59] under the Shuffle-LAGAN mode with the annotation of *Exbucklandia tonkinensis* plastome as a reference. To further investigate the presence of large-scale structural re-arrangements or genome structural variations among chloroplast genomes, Mauve [60] was employed to analyze 37 Hamamelidaceae species.

The boundary regions of the chloroplast genomes of the 37 Hamamelidaceae plants were visualized and compared using IRscope [61] (https://irscope.shinyapps.io/irapp/) to screen a macroscopic view of the chloroplast genome structure of the section and trace the size variation in the boundary regions of chloroplast genomes

among inverted repeat regions (IRs), small single copy region (SSC) and large single copy region (LSC).

Identification of microsatellites, repeats, and DNA barcodes of Hamamelidaceae

We used MIcroSAtellite (MISA) perl script [62] to identify chloroplast microsatellite markers (cpSSRs) within the cp genomes of the 37 Hamamelidaceae individuals, setting the parameters with thresholds of ten repeat units for mononucleotide SSRs, six repeat units for dinucleotide SSRs, five repeat units for tri-, tetra-, penta-, and hexa-nucleotide SSRs were employed. Family-targeted polymorphic SSRs across the 37 Hamamelidaceae individuals were further selected by the following three criteria: (1) cpSSRs located in the homologous regions; (2) cpSSRs possessed the same repeat units; and (3) the number of repeat units was variable.

REPuter [63] was used to count four types of scattered/dispersed sequence repeats (complement repeats, forward repeats, reverse repeats, and palindromic repeats) within 37 Hamamelidaceae chloroplast genomes, with the following parameter: the minimum repeat size of 50 bp and a hamming distance of eight.

DnaSP v.6.0 [64] (http://www.ub.edu/dnasp/) was used to find the DNA barcodes of Hamamelidaceae. Specifically, we calculated the nucleotide variability (π) of both coding regions and non-coding regions such as intergenic spacers and introns with aligned length > 200 bp and mutation site > 0 after sequence alignment via MAFFT v.7 [33] (https://mafft.cbrc.jp/alignment/software/) by DnaSP v.6.0. The calculated nucleotide variability values (π) were then plotted in R v.4.0.2 (https://www.r-project.org/).

Evolutionary mechanism and pan-genomic analysis of Hamamelidaceae

We employed the RSCU-Plot tool [65–68] (https://pcg-lab.shinyapps.io/RSCU-Plot/) for visualization and interpretation of codon usage patterns of Hamamelidaceae.

The synonymous (dS), non-synonymous (dN) nucleotide substitution rates and the dN/dS ratio (ω) of Hamamelidaceae chloroplast genomes were calculated using the CODEML program of PAML v.4.9 [69]. The pairwise dN and dS substitution rates between different taxa were calculated based on the custom selection model by setting CodonFreq prior as F3 × 4 model [70].

To estimate the pan-genome and core genome sizes of Hamamelidaceae, we utilized a custom Python script (pan_core_from_matrix.py) with a method based on random resampling [71]. All genomes were randomly shuffled and sequentially incorporated into the analysis to avoid the influence of genome inclusion order on the results. For each genome size, the pan-genome size and core genome size were calculated separately. Analysis was repeated 1,000 times, and the average values were used to construct the cumulative curves of the pan-genome and core genome [71].

Phylogenetic analysis of Hamameliadaceae

Based on two datasets—the complete chloroplast genome sequences and the shared chloroplast protein-coding genes (PCGs) among the sampled taxa—phylogenetic analyses were conducted using Maximum Likelihood (ML) and Bayesian Inference (BI) methods. The workflow was as follows:

(1) For the complete chloroplast genome sequences, multiple sequence alignment was performed using MAFFT v.7 [33]. The aligned sequences were then manually trimmed in Geneious Prime to remove poorly aligned and ambiguous regions. Substitution saturation of the aligned sequences was assessed using DAMBE v.5.3.74 [34]. Only sequences that showed no signs of substitution saturation were retained for subsequent phylogenetic reconstruction.
(2) For the chloroplast PCGs, gene sequences were first extracted for each species using Geneious Prime. Due to gene loss and/or the presence of pseudogenes identified during chloroplast genome annotation, only 76 PCGs were found to be shared among all 50 species. These 76 PCGs from each species were concatenated in Geneious Prime and aligned using MAFFT v.7 [33]. Poorly aligned and ambiguous regions were removed to generate a high-quality PCG data matrix for phylogenetic analysis. Substitution saturation of the aligned sequences was evaluated using DAMBE v.5.3.74 [34], and only sequences without saturation were retained for downstream phylogenetic reconstruction.
(3) ML analyses were performed using IQ-TREE v.1.6.12 [72]. The best-fit nucleotide substitution models were selected using ModelFinder [73] implemented in IQ-TREE, based on the Akaike Information Criterion (AIC). The best-fit model for the complete chloroplast genome dataset was TVM + F + R7, while that for the PCG dataset was GTR + F + R6. Node support was evaluated using 1,000 bootstrap (BS) replicates.
(4) For BI analyses, the best-fit substitution model was determined using MrModeltest v.2.3 [74], with GTR + I + G selected as the optimal model for both datasets. BI trees were constructed in MrBayes v.3.1.2 [75] under the GTR + I + G model (lset nst = 6, rates = invgamma; prset statefreqpr = dirichlet (1,1,1,1)). Two independent Markov chain Monte Carlo (MCMC) runs were conducted, each for 10,000,000 generations, with sampling every 1,000 generations. Convergence was assessed by

ensuring that the average standard deviation of split frequencies fell below 0.01. The first 25% of sampled trees were discarded as burn-in, and the remaining trees were used to construct a majority-rule consensus tree to estimate posterior probabilities (PP).

All resulting phylogenetic trees were visualized and edited using FigTree v.1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12870-026-08800-7.

Supplementary Material 1: Supplementary Figure 1. Sequence similarity plot of 37 Hamamelidaceae chloroplast genomes. *Exbucklandia tonkinensis* plastome is used as a reference genome. The Y-axis shows percentage of sequence similarity ranging from 50%–100%. The direction in which genes are transcribed is indicated by gray arrows. Non-coding sequences are depicted as pink bars, while exons are shown as purple bars. Variations in the genome are shown as white peaks.

Supplementary Material 2: Supplementary Figure 2. Synteny analysis of the chloroplast genomes of 37 Hamamelideceae species

Supplementary Material 3: Supplementary Figure 3. The relative synonymous codon usage (RSCU) values of chloroplast genomes of 37 Hamamelidaceae species

Supplementary Material 4: Supplementary Figure 4. The nonsynonymous (dN), synonymous (dS) substitution rates and dN/dS values of all plastid protein-coding genes (PCGs) of 37 Hamamelidaceae species

Supplementary Material 5: Supplementary Figure 5. The cumulative curve of the pan-genome and core genome of 37 Hamamelidaceae species

Supplementary Material 6: Supplementary Figure 6. Phylogeny of Hamamelidaceae based on the 76 concatenated protein-coding genes (PCGs) via ML and BI methods. The numbers at each node represent ML bootstrap support (BS) and BI posterior probability (PP), ML and BI tree show totally congruent topologies

Supplementary Material 7: Supplementary Table 1. Sampling information of 50 taxa used in the study of phylogenomics of Hamamelidaceae

Supplementary Material 8: Supplementary Table 2. Basic information statistics of 49 newly-sequenced chloroplast genomes of woody clade of Saxifragales

Supplementary Material 9:Supplementary Table 3. Statistics of SSR in chloroplast genome sequences of 37 Hamamelidaceae species

Supplementary Material 10: Supplementary Table 4. Statistics of scattered sequence repeats in chloroplast genome sequences of 37 Hamamelidaceae species

Supplementary Material 11: Supplementary Table 5. The list of nucleotide diversity (Pi) value of chloroplast gene fragment of 37 Hamamelidaceae species

Supplementary Material 12: Supplementary Table 6. The nonsynonymous (dN), synonymous (dS) substitution rates and dN/dS value of all plastid protein-coding genes (PCGs) of 37 Hamamelidaceae species

Acknowledgements
We thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript. We are grateful to Richard Ree (American Museum of Natural History), David Y. P. Tng (School for Field Studies), Jianhua Li (Hope College), Zhengwei Wang (Shanghai Chen Shan Botanical Garden), Yunfei Deng (South China Botanical Garden), Zulin Ning (South China Botanical Garden), Xingjian Liu (Nanjing Zhongshan Botanical Garden), Yiming Hu (Anhui Academy of Forestry), Lipeng Yu (Nature reserve of Longwangshan), Shanshan Zhu (Ningbo University), Wenming Li (Nanjing University) and En Shi (Nanjing University) for their great assistance with sampling.

Authors' contributions
Conceptualization, YZ.Z. and ZS.W.; methodology, YZ.Z. and YH.W.; validation, ZY.L., QX.C. and S.W., formal analysis, YZ.Z. and DYP.T., investigation, YZ.Z. and ZS.W.; resources, YZ.Z. and ZS.W.; data curation, YZ.Z. and ZS.W.; writing—original draft preparation, YZ.Z.; writing—review and editing, YH.W., QX.C., and ZY.L.; visualization, YZ.Z. and DYP.T.; supervision, YZ.Z. and YH.W.; project administration, YZ.Z. and ZS.W.; funding acquisition, YZ.Z. All authors have read and agreed to the submitted version of the manuscript.

Funding
This work was supported by the Jiangsu Funding Program for Excellent Postdoctoral Talent (2023ZB389).

Data availability
New sequenced and other published chloroplast genome sequences can be found in the GenBank (https://www.ncbi.nlm.nih.gov/genbank/), and the accession numbers are shown in **Supplementary Table 2**.

Declarations

Ethics approval and consent to participate
This study is not a clinical trial and the authors confirm that all methods complied with relevant institutional, national, and international guidelines and legislation. This article does not involve any endangered or protected species, and no specific permits are required for the collection of specimens.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 18 November 2025 / Accepted: 16 April 2026
Published online: 21 April 2026

References

1. Soltis DE, Mort ME, Latvis M, Mavrodiev EV, O'Meara BC, Soltis PC, et al. Phylogenetic relationships and character evolution analysis of Saxifragales using a supermatrix approach. Am J Bot. 2013;100:916–29.
2. Angiosperm Phylogeny Group IV. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Bot J Linn Soc. 2016;181:1–20.
3. Li JH. Molecular phylogenetics of Hamamelidaceae: evidence from DNA sequences of nuclear and chloroplast genomes. In: Sharma AK, Sharma A, editors. Plant genome: biodiversity and evolution. New Hampshire: Science Publisher; 2008. p. 227–50.
4. Xiang X, Xiang K, Ortiz RDC, Jabbour F, Wang W. Integrating palaeontological and molecular data uncovers multiple ancient and recent dispersals in the pantropical Hamamelidaceae. J Biogeogr. 2019;46:2622–31.
5. Wang S, Xie Y. China Species Red List. Beijing: Higher Education Press; 2004.
6. Liu Y, Liu K, Dong W, Dong S, Wang Y, Xu C, et al. Chloroplast genome evolution of Hamamelidaceae at subfamily level. Ecol Evol. 2025;15:e71141.
7. Li JH. Systematics of the Hamamelidaceae based on morphological and molecular evidence. PhD thesis. University of New Hampshire, Durham, NH, USA; 1997.
8. Gapinski A. Hamamelidaceae, Part 1: exploring the witch-hazels of the Arnold Arboretum. Arnoldia. 2014;72:2.
9. Meholic CA. Assessment of the Hamamelidaceae in global living collections. Master thesis. University of Delaware, Newark, Delaware, USA; 2019.
10. Zhang ZY, Lu AM. Hamamelidaceae: geographic distribution, fossil history and origin. Acta Phytotax Sin. 1995;33:313–39.
11. Endress PK. Aspects of evolutionary differentiation of the Hamamelidaceae and the lower Hamamelidae. Plant Syst Evol. 1989;162:193–211.

12. Gapinski A. Hamamelidaceae, Part 2: exploring the Witch-hazel relatives of the Arnold Arboretum. Arnoldia. 2015;72:4.
13. Maslova NP. Extinct and extant Platanaceae and Hamamelidaceae: morphology, systematics, and phylogeny. Paleontol J. 2003;37:5.
14. Shi SH, Chang HT, Chen YQ, Qu LH, Wen J. Phylogeny of the Hamamelidaceae based on the ITS sequences of nuclear ribosomal DNA. Biochem Syst Ecol. 1998;26:55–69.
15. Li JH, Bogle AL, Klein AS. Phylogenetic relationships of the Hamamelidaceae inferred from sequences of internal transcribed spacers (ITS) of nuclear ribosomal DNA. Am J Bot. 1999;86:1027–37.
16. Li JH, Bogle AL, Klein AS. Phylogenetic relationships in the Hamamelidaceae: evidence from the nucleotide sequences of the plastid gene matK. Plant Syst Evol. 1999;218:205–19.
17. Feng YX, Chen ZD, Wang XQ, Pan KY, Hong DY. A taxonomic revision of the *Loropetalum-Tetrathyrium* complex and its systematic position in the Hamamelidoideae based on morphology and ITS sequence data. Taxon. 1999;48:689–700.
18. Bobrov AV, Roslov MS, Romanov MS. Phylogenetic biogeography of Hamamelidaceae s. l. based on molecular data. Vestnik of Saint Petersburg University Earth Sciences. 2020;65:224–44.
19. Zhang YY, Li K, Feng Y, Li ZY, Hu YM, Yousefzadeh H, et al. Evolutionary legacy of the "living fossil" genus *Parrotia* (Hamamelidaceae): genomic insights into species divergence and polygenic adaptation. Plant J. 2025;123:e70367.
20. Dong WP, Liu YL, Xu C, Gao YW, Yuan QJ, Suo ZL, et al. Chloroplast phylogenomic insights into the evolution of *Distylium* (Hamamelidaceae). BMC Genomics. 2021;22:293.
21. Wang NJ, Chen SF, Xie L, Wang L, Feng YY, Lv T, et al. The complete chloroplast genomes of three Hamamelidaceae species: comparative and phylogenetic analyses. Ecol Evol. 2022;12:e8637.
22. Andrews S. Tree of the Year: *Parrotia*. Kew, UK: International Dendrology Society Press; 2007.
23. Li JH, Tredici PD. The Chinese *Parrotia*: a sibling species of the Persian *Parrotia*. Arnoldia. 2008;66:2–9.
24. Hao RM, Wei HT, Liu WG. Floral morphology of *Shaniodendron* (Hamamelidaceae) and its taxonomic significance. Plant Res. 1996;1:38–42.
25. Li JH, Bogle AL, Klein AS, Pan KY. Close relationship between *Shaniodendron* and *Parrotia* (Hamamelidaceae), evidence from ITS sequences of nuclear ribosomal DNA. J Syst Evol. 1997;35:481–93.
26. Birky CW. Uniparental inheritance of organelle genes. Curr Biol. 2008;18:692–5.
27. Brito PH, Edwards SV. Multilocus phylogeography and phylogenetics using sequence-based markers. Genetica. 2009;135:439–55.
28. Garrick RC, Bonatelli IA, Hyseni C, Morales A, Pelletier TA, Perez MF, et al. The evolution of phylogeographic data sets. Mol Ecol. 2015;24:1164–71.
29. Malé PGJ, Bardon L, Besnard G, Coissac E, Delsuc F, Engel J, et al. Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. Mol Ecol Resour. 2014;14:966–75.
30. Zhang SD, Jin JJ, Chen SY, Chase MW, Soltis DE, Li HT, et al. Diversification of Rosaceae since the Late Cretaceous based on plastid phylogenomics. New Phytol. 2017;214:1355–67.
31. Liu J, Zhang GF, Xue LX. Structural diversity and conservation implications of *Parrotia subaequalis* (Hamamelidaceae), a rare and endangered tree species in China. Nat Conserv. 2021;44:99–115.
32. Lu RS, Yang T, Chen Y, Wang SY, Cai MQ, Cameron KM, et al. Comparative plastome genomics and phylogenetic analyses of Liliaceae. Bot J Linn Soc. 2021;196:279–93.
33. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
34. Xia X. DAMBE5: a comprehensive software package for data analysis in molecular biology and evolution. Mol Biol Evol. 2013;30:1720–8.
35. Balakirev ES, Ayala FJ. Pseudogenes: are they "junk" or functional DNA? Annu Rev Genet. 2003;37:123–51.
36. Pei BK, Sisu C, Adam Frankish A, Howald C, Habegger L, Mu XM, et al. The gencode pseudogene resource. Genome Biol. 2012;13:R51.
37. Xie JB, Chen SS, Xu WJ, Zhao YY, Zhang DQ. Origination and function of plant pseudogenes. Plant Signal Behav. 2019;14:e1625698.
38. Mascagni F, Usai G, Cavallini A, Porceddu A. Structural characterization and duplication modes of pseudogenes in plants. Sci Rep. 2021;11:5292.
39. Kikuchi S, Bédard J, Hirano M, Hirabayashi Y, Oishi M, Imai M, et al. Uncovering the protein translocon at the chloroplast inner envelope membrane. Science. 2013;339:571–4.
40. De Vries J, Sousa FL, Bölter B, Soll J, Gould SB. YCF1: a green TIC? Plant Cell. 2015;27:1827–33.
41. Thomas F, Massenet O, Dorne AM, Briat JF, Mache R. Expression of the rpl23, rpl2 and rps19 genes in spinach chloroplasts. Nucleic Acids Res. 1988;16:2461–72.
42. Gao NN, Zhao ZL, Ni LH. Prospect: identification of medicinal plant based on chloroplast gene *ycf15*. Chinese Traditional and Herbal Drugs. 2017;48:3210–7.
43. Raubeson LA, Peery R, Chumley TW, Dziubek C, Fourcade HM, Boore JL, et al. Comparative chloroplast genomics: analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. BMC Genomics. 2007;8:174.
44. Shi C, Liu Y, Huang H, Xia EH, Zhang HB, Gao LZ. Contradiction between plastid gene transcription and function due to complex posttranscriptional splicing: an exemplary study of *ycf15* function and evolution in angiosperms. PLoS One. 2013;8:e59620.
45. Sun M, Ma Y, Yu J. From genomic fossils to functional elements: the evolving story of pseudogenes. Adv Genet. 2025;6:e00040.
46. Eriksen P, Oxelman B. Whole-gene positive selection, elevated synonymous substitution rates, duplication, and indel evolution of the chloroplast clpP1 gene. PLoS One. 2008;3:e1386.
47. Dobzhansky T. Nothing in biology makes sense except in the light of evolution. Am Biol Teach. 1973;35:125–9.
48. Soltis DE, Albert VA, Savolainen V, Hilu K, Qiu YL, Chase MW, et al. Genome-scale data, angiosperm relationships, and 'ending incongruence': a cautionary tale in phylogenetics. Trends Plant Sci. 2004;9:477–83.
49. Endress PK. Phylogenetic relationships in the Hamamelidoideae. In: Crane PR, Blackmore S, editors. Evolution, systematics and fossil history of the Hamamelidae. New York: Oxford University Press; 1989. p. 227–48.
50. Bogle AL, Philbrick CT. A generic atlas of Hamamelidaceous pollens. Contrib Gray Herb Harv. 1980;210:29–103.
51. Qi ZC, Yu Y, Liu X, Pais A, Ranney T, Whetten R, et al. Phylogenomics of polyploid *Fothergilla* (Hamamelidaceae) by RAD-tag based GBS-insights into species origin and effects of software pipelines. J Syst Evol. 2015;53:432–47.
52. Porebski S, Bailey LG, Baum BR. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. Plant Mol Biol Rep. 1997;15:8–15.
53. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
54. Jin JJ, Yu WB, Yang JB, Song Y, Li DZ. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol. 2020;21:241.
55. Patel RK, Jain M. NGS QC toolkit: a toolkit for quality control of next generation sequencing data. PLoS One. 2012;7:e30619.
56. Qu XJ, Moore MJ, Li DZ, Yi TS. PGA: a software package for rapid, accurate, and flexible batch annotation of plastomes. Plant Methods. 2019;15:50.
57. Schattner P, Brooks AN, Lowe TM. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. 2005;33:686–9.
58. Greiner S, Lehwark P, Bock R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Res. 2019;47:W59-64.
59. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 2004;32:W273-9.
60. Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004;14:1394–403.
61. Amiryousef A, Hyvonen J, Poczai P. The chloroplast genome sequence of bittersweet (*Solanum dulcamara*): plastid genome structure evolution in Solanaceae. PLoS One. 2018;13(4):e0196069.
62. Beier S, Thiel T, Münch T, Scholz U, Mascher M. MISA–web: a web server for microsatellite prediction. Bioinformatics. 2017;33:2583–5.
63. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R, et al. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 2001;29:4633–42.
64. Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, et al. DnaSP 6: DNA sequence polymorphism analysis of large datasets. Mol Biol Evol. 2017;34:3299–302.
65. Hejazi FA, Mohammadi P, Soorni A. Comparative chloroplast genomics of *Teucrium* species reveals genome evolution, phylogenetic relationships, and candidate molecular markers. Sci Rep. 2025;15:44318.

66. Diani GS, Soorni A. Plastome diversity and phylogenomic analysis of *Satureja* (Lamiaceae): uncovering evolutionary patterns and diagnostic markers. Planta. 2025;262:149.
67. Akrami AM, Meratian ES, Soorni A. Decoding the chloroplast genomes of five Iranian *Salvia* species: insights into genomic structure, phylogenetic relationships, and molecular marker development. BMC Genomics. 2025;26:545.
68. Soorni A, Golchini MM. Complete chloroplast genome of *Mentha aquatica* reveals hypervariable regions and resolves phylogenetic position within the genus *Mentha*. Mol Biol Rep. 2025;52:677.
69. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24:1586–91.
70. Yang ZH, Nielsen R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol. 2002;19:908–17.
71 Zekic T, Holley G, Stoye J. Pan-genome storage and analysis techniques in: Setubal J, Stoye J, Stadler P, editors. Comparative genomics. Methods Mol Biol. 2018;1704:29–53. New York, NY: Humana Press.
72. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32:268–74.
73. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14:587–9.
74. Posada D, Buckley TR. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst Biol. 2004;53:793–808.
75. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 2003;19:1572–4.

Publisher's Note