Plastid phylogenomic insights into the evolution of *Distylium* (Hamamelidaceae)

Wenpan Dong
Beijing Forestry University
Yanlei Liu
Institute of Botany Chinese Academy of Sciences
Chao Xu
Institute of Botany Chinese Academy of Sciences
Yongwei Gao
Beijing Forestry University
Zhixiang Zhang
Beijing Forestry University
Qingjun Yuan
China Academy of Chinese Medical Sciences
Zhili Suo
Institute of Botany Chinese Academy of Sciences
Jiahui Sun (✉ sunjh_2010@sina.com)
China Academy of Chinese Medical Sciences

Research article

Keywords: Distylium, Hamamelidaceae, species identification, cpDNA marker, phylogenomics

Posted Date: December 2nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-119007/v1

Version of Record: A version of this preprint was published at BMC Genomics on April 22nd, 2021. See the published version at https://doi.org/10.1186/s12864-021-07590-6.

Abstract

Background: Most *Distylium* species are endangered. *Distylium* species mostly display homoplasy in their flowers and fruits, and are classified primarily based on leaf morphology. However, leaf size, shape, and serration vary tremendously making it difficult to use those characters to identify most species and a significant challenge to address the taxonomy of *Distylium.* To infer robust relationships and identify variable markers to identify *Distylium* species, we sequenced most of the *Distylium* species chloroplast genome.

Results: The *Distylium* chloroplast genome size was 159,041–159,127 bp and encoded 80 protein-coding, 30 transfer RNAs, and 4 ribosomal RNA genes. There was a conserved gene order displayed and a typical quadripartite structure. Phylogenomic analysis based on whole chloroplast genome sequences yielded a highly resolved phylogenetic tree and formed a monophyletic group containing four *Distylium* clades. A dating analysis suggested that *Distylium* originated in the Oligocene (34.39 Ma) and diversified within approximately 1 Ma. The evidence shows that *Distylium* is a rapidly radiating group. Four highly variable markers, such as *matK-trnK*, *ndhC-trnV*, *ycf1*, and *trnT-trnL*, and 74 polymorphic simple sequence repeats were discovered in the *Distylium* plastomes.

Conclusions: The plastome sequences had sufficient polymorphic information to resolve phylogenetic relationships and identify species accurately.

Background

Distylium Sieb. et Zucc is a genus of flowering plants in the tribe Fothergilleae of the family Hamamelidaceae, which is endemic to Asia. Fifteen species have been reported in *Distylium* worldwide, with 12 species occurring in China (*D. chinense*, 2n = 24). Additionally, two species are found in Japan, one of which is found also in China, and one species each in Malaysia and India. They are evergreen shrubs or small trees that grow mostly in subtropical evergreen forests.

This genus has been introduced as a cultivar and thrives in warm temperate and subtropical climates in Europe and the United States. *Distylium*, with dense branches and deep evergreen leaves, a neat tree shape, small red flowers in spring, good soundproof effects, and strong resistance to smoke and dust and various toxic gases (e.g., sulfur dioxide and chlorine), are suitable as greening and ornamental plants in cities, and industrial and mining areas. They are commonly cultivated in urban gardens in the Yangtze River basin of China. Some species, such as *D. chinense*, are used to stabilize solid earth embankments because of their robust root system, flooding tolerance, and resistance to sand burial soaks [1, 2].

Most *Distylium* species are endangered. According to the threatened species list of China's higher plants [3], two species are Critically Endangered species (*D. macrophyllum* and *D. tsiangii*), two are Endangered species (*D. chinense* and *D. gracile*), and two species are Vulnerable (*D. chungii* and *D. elaeagnoides*). Some *Distylium* species are narrowly distributed, such as *D. lepidotum*, which is endemic to the Ogasawara (Bonin) Islands, located in the northwestern Pacific approximately 1,000 km south of Tokyo [4]. *D. tsiangii* is only located in Dushan and Bazai counties of Guizhou Province.

Distylium species lack significant differences in the morphology of their flowers and fruits, and are classified primarily based on leaf morphology. However, leaf size, shape, and serration vary tremendously and are difficult characters to use in most cases. For example, the range of leaf variation in *D. buxifolium* is very striking [5]. This variability has led to a proposed number of new species, which have been reduced to synonymy, as more material has been found to link extreme forms [5]. Due to the insufficient number of morphological diagnostic characters and highly polymorphic traits, taxonomic classification of *Distylium* species has been unclear. However, chloroplast genome markers, such as *atpB*, *atpB-rbcL*, *matK*, *rbcL*, *trnH-psbA*, and *trnL-F*, and the internal transcribed spacer (ITS) has enabled molecular phylogenetic analyses of several *Distylium* species [6–9]. However, those markers have lower divergence among *Distylium* species; moreover, no study has inferred the phylogeny of this genus.

Whole chloroplast genome sequences have been widely used to infer phylogenetic relationships at different taxonomic levels, and provide an effective genetic resource for resolving complex evolutionary relationships and identifying ambiguous species. With the development of sequencing methods, complete chloroplast genome sequences are now available at low cost, extending gene-based phylogenetics to genome-based phylogenomics [10–12], extending gene-based species identification to genome-based super DNA barcoding [13, 14], and making it easier to study evolutionary events in plant species [15].

In this study, we specifically aimed to (1) develop and screen appropriate intrageneric markers in the chloroplast genome to establish DNA barcodes for *Distylium*; (2) estimate the effectiveness of a whole chloroplast genome data set in resolving the relationships within this radiating lineage; (3) estimate the divergence time of *Distylium*.

Results

Basic characteristics of the *Distylium* plastomes

The complete chloroplast genomes of the 12 newly sequenced *Distylium* species ranged in length from 159,041 bp (*D. lepidoium*) to 159,127 bp (*D. gracile*) (Table 1). The *Distylium* chloroplast genomes had a quadripartite structure typical of most angiosperm species, including large single copy (LSC) and small single copy (SSC) regions separated by two inverted repeat (IRa and IRb) regions (Fig. 1). The LSC regions ranged from 87,825 bp (*D. pingpienense*) to 87,863 bp (*D. racemosum*), the SSC regions varied between 18,770 bp (*D. dunnianum*) and 18,796 bp (*D. lepidoium*), and the IR regions ranged from 26,225 bp (*D. elaeagnoides*) to 26,241 bp (*D. dunnianum*). The GC content of the chloroplast genome sequences was 38.0%. A total of 114 unique genes was detected in the chloroplast genomes of the 11 *Distylium* species, including 80 protein coding genes, 30 tRNA genes, and 4 rRNA genes, and the gene order was highly conserved (Fig. 1 and Table 1).

Table 1
The basic plastomes information of 12 *Distylium* samples.

	Nucleatide length (bp)				**Nmuber of genes**			**GC%**				
Species	Total	LSC	SSC	IR	Protein	tRNA	rRNA	Total	LSC	SSC	IR	Genbank accession number
D. buxifolium	159,084	87,828	18,790	26,233	80	30	4	38.0	36.2	32.5	43.1	MW248115
D. chinese	159,087	87,830	18,791	26,233	80	30	4	38.0	36.2	32.5	43.1	MW248112
D. cuspidatum	159,068	87,848	18,784	26,218	80	30	4	38.0	36.2	32.4	43.1	MW248117
D. dunnianum	159,097	87,845	18,770	26,241	80	30	4	38.0	36.2	32.5	43.1	MW248109
D. elaeagnoides	159,094	87,857	18,787	26,225	80	30	4	38.0	36.2	32.5	43.1	MW248120
D. gracile	159,127	87,854	18,793	26,240	80	30	4	38.0	36.2	32.5	43.0	MW248116
D. lepidoium	159,041	87,831	18,796	26,205	80	30	4	38.0	36.2	32.5	43.1	MW248118
D. lepidoium	159,042	87,832	18,796	26,205	80	30	4	38.0	36.2	32.5	43.1	MW248119
D. macrophyllum	159,095	87,847	18,788	26,230	80	30	4	38.0	36.2	32.5	43.1	MW248111
D. myricoides	159,093	87,847	18,780	26,233	80	30	4	38.0	36.2	32.5	43.1	MW248110
D. pingpienense	159,081	87,825	18,790	26,233	80	30	4	38.0	36.2	32.5	43.1	MW248114
D. racemosum	159,107	87,863	18,782	26,231	80	30	4	38.0	36.2	32.5	43.1	MW248113

Repetitive sequences

A total of 801 SSRs were identified across the chloroplast genomes of the 11 *Distylium* species (Fig. 2 and Table S2). The number of SSRs per species ranged from 70 (*D. dunnianum*) to 78 (*D. gracile*). The majority of the SSRs were mononucleotide repeats (78.65%), followed by dinucleotide (8.61%) and tetranucleotide (5.87%) repeats. There were no hexanucleotide repeats in the *Distylium* plastomes. The SSR A and T motifs were the most frequent. SSRs were particularly rich in AT in the *Distylium* plastomes. Among those SSRs, most were located in the LSC/SSC regions (94.01%).

A total of 96 unique SSRs and 74 SSRs were polymorphic across the 11 *Distylium* species. All polymorphic SSRs were located in the single copy regions, except two SSRs (Table 2). The mononucleotide repeat units A and T were also the most frequent polymorphic

SSRs.

Table 2
Polymorphism of SSRs in *Distylium* plastomes.

Regions/SSR unit	Overall	Polymorphic	Monomorphic
LSC	80	60	20
IR	2	2	0
SSC	14	12	2
A	30	30	0
T	42	36	6
C	3	0	3
G	1	1	0
TA	4	2	2
AT	3	1	2
TC	1	0	1
TTA	2	1	1
ATA	1	0	1
TAT	1	1	0
GAA	1	0	1
AAAT	1	0	1
ATAC	1	0	1
GAAA	1	1	0
TATTT	1	0	1
TGAA	1	0	1
TTCT	1	1	0
TTCTA	1	0	1
Total	96	74	22

Indel variations

A total of 76 indels were discovered in the *Distylium* plastomes, including 59 normal indels and 17 repeat indels. Most of the indels (72.37%, 55 times) were located in the spacer regions, 15.79% (12 times) of indels occurred in the exons, and 11.84% (nine times) were found in the introns (Fig. 3). The *TrnT-trnL* spacer had five indels, followed by *ndhC-trnV* (3 indels). The size of the normal indels ranged from 1 to 13 bp, with 8 bp and 9 bp length indels being the most common. The largest indel (13 bp) was located in the *trnC-petN* spacer and was a deletion in *D. macrophyllum*. The second largest indel was in the *ycf1* exon of 12 bp length and was an insert in the two *D. lepidoium* samples. The length of the repeat indels ranged from 2 to 16 bp. The largest repeat indel occurred in the *rpl20-rps12* spacer and the second largest repeat indel was located in the *rps7-trnV* spacer.

Variation in the plastomes and molecular markers for *Distylium* species

The entire chloroplast genome of the 11 *Distylium* species was 159,360 bp in length, including 298 polymorphic sites and 115 parsimony informative sites (Table 3). The overall nucleotide diversity (π) was 0.00045; however, each region of the chloroplast genome revealed different nucleotide diversity; the SSC exhibited the highest π value (0.00089) and the IR had the lowest π value (0.00006). All species had a unique chloroplast haplotype, except the IR regions. The number of nucleotide substitutions among the 11 species varied from 7 to 109, and the p-distance varied from 0.0004 to 0.0069. The lowest divergence was between *D. buxifolium* and *D. chinese*, and the largest sequence divergence was observed between *D. chinese* and *D. lepidoium*.

Table 3
Sequences divergence of *Distylium* plastomes.

Regions	Alignment length (bp)	Number of variable sites			Nucleotide polymorphism	
		Polymorphic	Singleton	Parsimony informative	Nucleotide diversity	Haplotypes
LSC	88,033	210	125	85	0.00059	11
SSC	18,825	74	48	26	0.00089	11
IR	26,251	7	5	2	0.00006	7
Whole plastomes	159,360	298	183	115	0.00045	11

The π value ranged from 0 to 0.0027 in an 800-bp sliding window size. In total, four peaks with π values > 0.002 were identified in the chloroplast genome (Fig. 4). Those regions included *matK-trnK*, *ndhC-trnV*, *ycf1*, and *trnT-trnL*. Three intergenic regions (*matK-trnK*, *ndhC-trnV*, and *trnT-trnL*) were located in the LSC region, and the *ycf1* coding region was in the SSC region.

We tested the variability in the hypervariable markers by comparing the chloroplast genome and the three universal DNA barcodes (*matK*, *rbcL*, and *trnH-psbA*). The variable information is shown in Table 4. The intergenic spacer marker *trnH-psbA* was 367 bp, including two variable sites and no parsimony informative sites. The *rbcL* and *matK* genes were 1,428 bp with three variable and three informative sites, and 1,515 bp with only one variable and no informative sites. Combined with the three universal markers, the aligned length was 3,310 bp, with six variable sites and three informative sites. The mean distance was 0.00045. The species identification analyses showed that the universal DNA barcodes had less discriminatory power; they had only four haplotypes when combined with the three markers, and the NJ tree had lower resolution and most of the samples were not distinguished (Table 4 and Fig. 5).

Table 4
Variability of the four highly mutation hotspot regions and the universal chloroplast DNA barcodes in *Distylium*.

Markers	Length (bp)	Polymorphic sites	Parsimony information sites	Mean distance	Nucleotide diversity	Number of haplotype
matK-trnK	827	8	3	0.00228	0.00227	7
trnT-trnL	1,170	9	4	0.00184	0.00173	7
ndhC-trnV	961	6	4	0.00197	0.00198	7
ycf1	2,306	20	5	0.00179	0.00179	9
Combination four variable markers	5,264	43	16	0.00191	0.00197	11
trnH-psbA	367	2	0	0.00084	0.00084	2
matK	1,515	1	0	0.00010	0.00010	2
rbcL	1,428	3	3	0.00072	0.00072	3
Combination three universal markers	3,310	6	3	0.00045	0.00045	4

The four hypervariable markers ranged from 827 bp (*matK-trnK*) to 2,306 bp (*ycf1*) in length. The *ycf1* gene had the greatest number of variable sites (20 sites) followed by *trnT-trnL* (9 sites), *matK-trnK* (8 sites), and *ndhC-trrnV* had the fewest (6 sites). Combining the four hypervariable markers, there were 43 variable sites and 16 parsimony informative sites that produced the most current identification (Table 4). The identified hypervariable markers had higher resolution compared with the tree universal markers, based on the NJ tree (Fig. 5).

Phylogenetic inference

Using the complete chloroplast genome sequences, we inferred the phylogenetic relationships among the 24 Hamamelidaceae samples. The topology of the ML and BI trees was nearly identical (Fig. 6). All *Distylium* species formed a monophyletic clade that was sister to *Parrotia* within Fothergilleae. *Distylium* had a shortened branch on the phylogenetic tree, indicating low divergence among *Distylium* species. Four clades were reconstructed in *Distylium* with a 100% bootstrap value. Clade I included the basal species *D. lepidoium*. Clade II included only *D. myricoides*. Clade III included only *D. macrophyllum*. Clade IV included the most advanced eight species, i.e., *D. buxifolium*, *D. chinense*, *D. pingienense*, *D. cuspidatum*, *D. dunnianum*, *D. gracile*, *D. elaeagoides*, and *D. racemosum* (Fig. 6).

Estimate of divergence time

Divergence time estimates suggested that Hamamelioideae diverged from Hamamelidaceae about 99.38 Ma (95% HPD: 90.71–105.44 Ma) during the Cenomanian of the Upper Cretaceous (Fig. 7). The stem note of Fothergilleae was dated to 88.87 Ma (95% HPD: 97–91.18 Ma). The stem date for *Distylium* was estimated to be 34.39 Ma (95% HPD: 29.99–39.03 Ma) in the Oligocene and the *Distylium* crown date was 5.39 Ma (95%HPD: 0.82–12.3 Ma) in the Pliocene. Diversification with this genus occurred over a short time period of approximately 1 Ma.

Discussion

The genera *Distyliopsis*, *Distylium*, *Fothergilla*, *Parrotia*, *Parrotiopsis*, *Shaniodendron*, and *Sycopsis* occur in in the tribe Fothergilleae of the subfamily Hamamedoideae [9]. According to the phylogenetic relationships based on the several chloroplast and nuclear ITS genes [6, 8], *Distylium* is sister to *Distyliopsis* [9]. This is the first use of molecular data to infer the *Distylium* phylogeny. The *Distylium* genus formed a well-defined monophyletic group according to the chloroplast genome data (Fig. 6). Moreover, the phylogenetic tree possessed a series of short internodes within *Distylium* and most species diversified < 1 Ma (Fig. 6), suggesting that this genus has undergone rapid radiation. *D. lepidoium* was at the base of the genus. This species was first described in 1918 and is endemic to the Ogasawara Islands [4]. *D. myricoides* formed a monotypic clade and is distributed in eastern and southeastern China. According to the morphological characteristics, *D. myricoides* resembles *D. buxijolium* most closely, from which it may be distinguished by its larger leaves [5]. However, this relationship was not supported by the present study. *D. buxijolium* and *D. chinense* were sister species and formed a group supported by morphological characteristics [5]. In this study, the chloroplast genome data provided information to infer the phylogeny of *Distylium.* However, due to rapid radiation, sampling of additional individuals from each species and extending more nuclear genes would provide additional evidence of the evolutionary history of *Distylium.*

Most *Distylium* species are rare and endangered; thus, the development of rapid and easily accessible species identification methods is essential. The variations in the morphological characteristics between species were continuous and uninterrupted. Therefore, it was difficult to distinguish species using morphological characteristics. DNA barcoding offers an opportunity to identify *Distylium* species. *rbcL* and *matK* are the two core DNA barcodes in plants. However, many studies have shown that these two markers have lower species identification power [16, 17]. Our study also showed that *rbcL* and *matK* or a combination of the two markers failed to discriminate *Distylium* species (Fig. 5), explaining the low resolution in previous studies and highlighting the importance of developing highly divergent markers.

Some studies have indicated that mutations are not random and are clustered as “mutation hotspots” or “highly variable regions” [10, 16, 18]. In this study, we compared the whole chloroplast genomes and identified the mutation hotspots in *Distylium* (Fig. 3). Four variable loci (*matK-trnK*, *ndhC-trnV*, *ycf1*, and *trnT-trnL*) were discovered. *trnT-trnL* has been frequently used in plant phylogeny [19]. *NdhC-trnV* and *ycf1* are considered divergence hotspots in angiosperms based on our previous research [16]. *NdhC-trnV* has been less used in plant phylogeny and species identification and is prone to large indels [20]. The coding region of the *ycf1* locus is the most divergent marker in most groups, and has been suggested as the main plant DNA barcode [17]. *matK-trnK* is located in the LSC region, and this locus is used less frequently in evolutionary biology. Some lineages have the Ploy T structure [21]. Therefore, the lineage-specific, highly variable markers developed in this study will facilitate further phylogenetic reconstruction and DNA barcoding of rare and endangered *Distylium* species.

Conclusions

In this study, we report 10 newly sequenced chloroplast genomes of *Distylium* species. The overall genomic structure, including the gene number and gene order, was well-conserved. The phylogeny and divergence time analyses based on the plastome sequences showed that *Distylium* was a rapidly radiating group and most speciation events occurred < 1 Ma. A comparison of sequence divergence across the *Distylium* plastomes revealed that *matK-trnK*, *ndhC-trnV*, *ycf1*, and *trnT-trnL* were mutation hotspot regions. Overall, our study demonstrated that plastome sequences can be used to improve phylogenetic resolution and species discrimination. Extended sampling and additional nuclear markers are absolutely necessary in further studies.

Methods

Plant material and DNA extraction

A total of 12 individual samples representing 11 *Distylium* species were sampled from the Plant DNA Bank of China at the Institute of Botany, Chinese Academy of Sciences. All samples were identified based on morphological characters. The details of the plant samples are presented in Table S1. Total genomic DNA was extracted from the leaf tissues of herbarium specimens of this genus following the modified CTAB DNA extraction protocol [22].

Sequence, chloroplast genome assembly, and annotation

The total DNA was constructed using 350-bp insert libraries according to the manufacturer's instructions, which was then used for sequencing. Paired-end sequencing was performed on an Illumina HiSeq X-ten at Novogene (Tianjin, China), yielding approximately 4 Gb of high-quality 150-bp paired-end reads per sample.

The raw reads obtained from Novogene were filtered using Trimmomatic 0.39 [23] with the following parameters: LEADING = 20, TRAILING = 20, SLIDING WINDOW = 4:15, MIN LEN = 36, and AVG QUAL = 20. High-quality reads were assembled *de novo* using the SPAdes 3.6.1 program [24]. The chloroplast genome sequence contigs were selected from the initial assembled reads in SPAdes by performing a BLAST search using several related Hamamelidaceae chloroplast genome sequences as references. The chloroplast genome sequence contigs were further assembled using Sequencher 5.4.5. All plastid assemblies were annotated in Plann [25] using *D. macrophyllum* (GenBank Accession number: MN729500) as the reference, and missing or incorrect genes were checked in Sequin. A circular diagram for the chloroplast genome was generated using OGDRAW [26]. All chloroplast genomes assembled in this study have been deposited in GenBank under Accession numbers MW248109 - MW248120.

Microstructural mutation events

The Perl script microsatellite identification tool (MISA, http://pgrc.ipk-gatersleben.de/misa/misa.html) was used to identify the microsatellite regions of the chloroplast genome with the parameters set to 10 (repeat units ≥ 10) for mononucleotide simple sequence repeats (SSRs), 6 (repeat units ≥ 6) for dinucleotides, 5 (repeat units ≥ 5) for trinucleotides, 4 (repeat units ≥ 4) for tetranucleotides, and 3 (repeat units ≥ 3) for pentanucleotides and hexanucleotides.

The chloroplast genomes sequences were aligned using MAFFT [27] manually examined, and adjusted. Based on the aligned sequence matrix, the indels were manually checked and divided into categories of repeat indels and normal indels, according to Dong et al. [15]. *D. dunnianum* was used as the reference to determine the size and position of the indel events.

Sequence divergence analysis

Single nucleotide substitutions and the genetic p-distances were calculated using MEGA 7.0 [28] based on the aligned chloroplast genome sequences. To assess sequence divergence and to explore highly variable chloroplast markers, nucleotide diversity (π) was calculated by sliding window analysis using DnaSP v6 [29] with a widow size of 600 bp and a step size of 100 bp.

Nucleotide diversity and the number of haplotypes were used to assess marker variable for all barcodes (hype-variable markers and the universal plant DNA barcodes, such as *rbcL*, *matK*, and *trnH-psbA*). The tree-based method was utilized to calculate discrimination power. A neighbor-joining (NJ) tree was prepared in PAUP using the K2p distance.

Phylogenetic analyses

To elucidate the phylogenetic positions of *Distylium* within Hamamelidaceae and the interspecific phylogenetic relationships within *Distylium*, multiple alignments were performed using the whole chloroplast genome of 24 Hamamelidaceae samples representing 11 genera, including *Cercidiphyllum japonicum, Daphniphyllum oldhamii*, and *Liquidambar formosana* as outgroups. The Hamamelidaceae chloroplast genomes were aligned using MAFFT, and ambiguous alignment regions were trimmed with Gblocks 0.91b [30]. The maximum-likelihood (ML) analysis was run with RAxML-NG [31] with the best-fit model from ModelFinder [32]. Branch support was assessed by fast bootstrap methodology using non-parametric bootstrapping and 500 ML pseudo-replicates.

Mrbayes v3.2 [33] was used to infer the Bayesian inference (BI) tree. The BI analysis was run for 20 million generations, in which a tree was sampled every 1,000 generations. Two independent Markov Chain Monte Carlo (MCMC) analyses were performed and each chain started with a random tree. The first 25% of the sampled trees was discarded as burn-in, while the remaining trees were constructed in a majority-rule consensus tree to estimate posterior probabilities.

Molecular clock dating

We used BEAST v2.5.1 [34] to estimate the divergence times of Hamamelidaceae using three priors based on the complete plastome sequences. Based on the average value obtained by Xiang et al. [9] in a calibrated analysis, three priors were used: (i) the average age of the most recent common ancestor (TMRCA) of Hamamelidaceae (the root of the tree) was 108 Ma; (ii) the crown age of Hamamelideae/Fothergilleae was 89 Ma; and (iii) the crown age of Mytilarioideae was 58.3 Ma. Each secondary prior was placed under a normal distribution with a standard deviation of 1.

The GTR nucleotide substitution model and the prior tree Yule model were selected with the uncorrelated lognormal distribution relaxed molecular clock model. The MCMC run had a chain length of 400,000,000 generations with sampling every 10,000 generations. The stationary phase was examined through Tracer 1.6 [35] to evaluate convergence and to ensure sufficient and effective sample size for all parameters surpassing 200. A burn-in of 10% generations was discarded, and TreeAnnotator v2.4.7 was used to produce a maximum clade credibility tree.

Abbreviations

BI: Bayesian Inference; bp: base pairs; Gb: Gigabases; LSC: Long single copy; Ma: million years ago; MCMC: Markov chain Monte Carlo; ML: Maximum likelihood; NCBI: National Center for Biotechnology Information; NGS: Next generation sequencing; π: nucleotide diversity; rRNA: ribosomal RNA; SSC: Short single copy; SSR: Simple sequence repeat; tRNA: transfer RNA.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was funded by the National Key Research and Development Program of China (2017YFC1703700 & 2017YFC1703704), the National Natural Science Foundation of China (NSFC: 81891010 & 81891014).

Authors’ contributions

Wenpan Dong: Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft, Writing - review & editing. Yanlei Liu: Methodology, Data curation. Chao Xu: Resources, Writing - original draft. Yongwei Gao: Methodology, Software. Zhixiang Zhang: Resources. Qingjun Yuan: Resources. Zhili Suo: Resources. Writing - original draft, Jiahui Sun: Supervision, Funding acquisition.

Acknowledgments

We would like to thank the Plant DNA Bank of China in the Institute of Botany, Chinese Academy of Sciences for providing materials.

Data availability

The chloroplast genome of *Distylium* under study is deposited in the GenBank database under the following accession numbers: MW248109 - MW248120.

References

1. Liu Z, Cheng R, Xiao W, Guo Q, Wang N: **Effect of Off-Season Flooding on Growth, Photosynthesis, Carbohydrate Partitioning, and Nutrient Uptake in Distylium chinense**. *PLOS ONE* 2014, **9**(9):e107636.
2. Xiang L, Li X-L, Wang X-S, Yang J, Lv K, Xiong Z-Q, Chen F-Q, Huang C-M: **Genetic diversity and population structure of Distylium chinense revealed by ISSR and SRAP analysis in the Three Gorges Reservoir Region of the Yangtze River, China**. *Global Ecology and Conservation* 2020, **21**:e00805.
3. Qin H, Yang Y, Dong S, He Q, Jia Y, Zhao L, Yu S, Liu H, Liu B, Yan Y *et al*: **Threatened Species List of China's Higher Plants**. *Biodiversity Science* 2017, **25**(7):696-744.
4. Yagi H, Xu J, Moriguchi N, Miyagi R, Moritsuka E, Sato E, Sugai K, Setsuko S, Torimaru T, Yamamoto S-i *et al*: **Population genetic analysis of two species of Distylium: D. racemosum growing in East Asian evergreen broad-leaved forests and D. lepidotum endemic to the Ogasawara (Bonin) Islands**. *Tree Genetics & Genomes* 2019, **15**(6):77.
5. Walker EH: **A revision of Distylium and Sycopsis (Hamamelidaceae)**. *Journal of the Arnold Arboretum* 1944, **25**(3):319-341.
6. Shi S, Chang HT, Chen Y, Qu L, Wen J: **Phylogeny of the Hamamelidaceae based on the ITS sequences of nuclear ribosomal DNA**. *Biochemical Systematics and Ecology* 1998, **26**(1):55-69.
7. Li J, Bogle AL, Klein AS: **Phylogenetic relationships in the Hamamelidaceae: Evidence from the nucleotide sequences of the plastid gene matK**. *Plant Syst Evol* 1999, **218**(3):205-219.
8. Li J, 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**(7):1027-1037.
9. 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**(11):2622-2631.
10. Dong W, Xu C, Li W, Xie X, Lu Y, Liu Y, Jin X, Suo Z: **Phylogenetic resolution in *Juglans* based on complete chloroplast genomes and nuclear DNA sequences**. *Frontiers in Plant Science* 2017, **8**:1148.
11. Dong W, Xu C, Wu P, Cheng T, Yu J, Zhou S, Hong D-Y: **Resolving the systematic positions of enigmatic taxa: Manipulating the chloroplast genome data of Saxifragales**. *Mol Phylogenet Evol* 2018, **126**:321-330.
12. Guo L, Guo S, Xu J, He L, Carlson JE, Hou X: **Phylogenetic analysis based on chloroplast genome uncover evolutionary relationship of all the nine species and six cultivars of tree peony**. *Industrial Crops and Products* 2020, **153**:112567.
13. Chen X, Zhou J, Cui Y, Wang Y, Duan B, Yao H: **Identification of Ligularia Herbs Using the Complete Chloroplast Genome as a Super-Barcode**. *Frontiers in Pharmacology* 2018, **9**(695).
14. Krawczyk K, Nobis M, Myszczynski K, Klichowska E, Sawicki J: **Plastid super-barcodes as a tool for species discrimination in feather grasses (Poaceae: *Stipa*)**. *Sci Rep* 2018, **8**(1):1924.
15. Dong W, Xu C, Wen J, Zhou S: **Evolutionary directions of single nucleotide substitutions and structural mutations in the chloroplast genomes of the family Calycanthaceae**. *BMC Evol Biol* 2020, **20**(1):96.
16. Dong W, Liu J, Yu J, Wang L, Zhou S: **Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding**. *PLOS ONE* 2012, **7**(4):e35071.
17. Dong W, Xu C, Li C, Sun J, Zuo Y, Shi S, Cheng T, Guo J, Zhou S: **ycf1, the most promising plastid DNA barcode of land plants**. *Sci Rep* 2015, **5**:8348.
18. Li W, Liu Y, Yang Y, Xie X, Lu Y, Yang Z, Jin X, Dong W, Suo Z: **Interspecific chloroplast genome sequence diversity and genomic resources in *Diospyros***. *BMC Plant Biol* 2018, **18**(1):210.
19. Hamzeh M, Dayanandan S: **Phylogeny of Populus (Salicaceae) based on nucleotide sequences of chloroplast TRNT-TRNF region and nuclear rDNA**. *Am J Bot* 2004, **91**(9):1398-1408.

20. Shaw J, Lickey EB, Schilling EE, Small RL: **Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: The tortoise and the hare III**. *Am J Bot* 2007, **94**(3):275-288.
21. Wheeler GL, McGlaughlin ME, Wallace LE: **Variable length chloroplast markers for population genetic studies in Acmispon (Fabaceae)**. *Am J Bot* 2012, **99**(10):e408-e410.
22. Li J, Wang S, Jing Y, Wang L, Zhou S: **A modified CTAB protocol for plant DNA extraction**. *Chin Bull Bot* 2013, **48**(1):72-78.
23. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data**. *Bioinformatics* 2014, **30**(15):2114-2120.
24. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD *et al*: **SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing**. *J Comput Biol* 2012, **19**(5):455-477.
25. Huang DI, Cronk QCB: **Plann: A command-line application for annotating plastome sequences**. *Applications in Plant Sciences* 2015, **3**(8):1500026.
26. 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**(W1):W59-W64.
27. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability**. *Mol Biol Evol* 2013, **30**(4):772-780.
28. Kumar S, Stecher G, Tamura K: **MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets**. *Mol Biol Evol* 2016, **33**(7):1870-1874.
29. Rozas J, Ferrer-Mata A, Sanchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sanchez-Gracia A: **DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets**. *Mol Biol Evol* 2017, **34**(12):3299-3302.
30. Castresana J: **GBLOCKS: selection of conserved blocks from multiple alignments for their use in phylogenetic analysis**. *EMBL* 2002.
31. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A: **RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference**. *Bioinformatics* 2019, **35**(21):4453-4455.
32. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS: **ModelFinder: fast model selection for accurate phylogenetic estimates**. *Nat Methods* 2017, **14**(6):587-589.
33. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP: **MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space**. *Syst Biol* 2012, **61**(3):539-542.
34. Bouckaert R, Heled J, Kuhnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ: **BEAST 2: a software platform for Bayesian evolutionary analysis**. *PLoS Comp Biol* 2014, **10**(4):e1003537.
35. Rambaut A, Suchard M, Xie D, Drummond A: **Tracer v1. 6**. In.; 2014: Available from http://beast.bio.ed.ac.uk/Tracer.

Figures

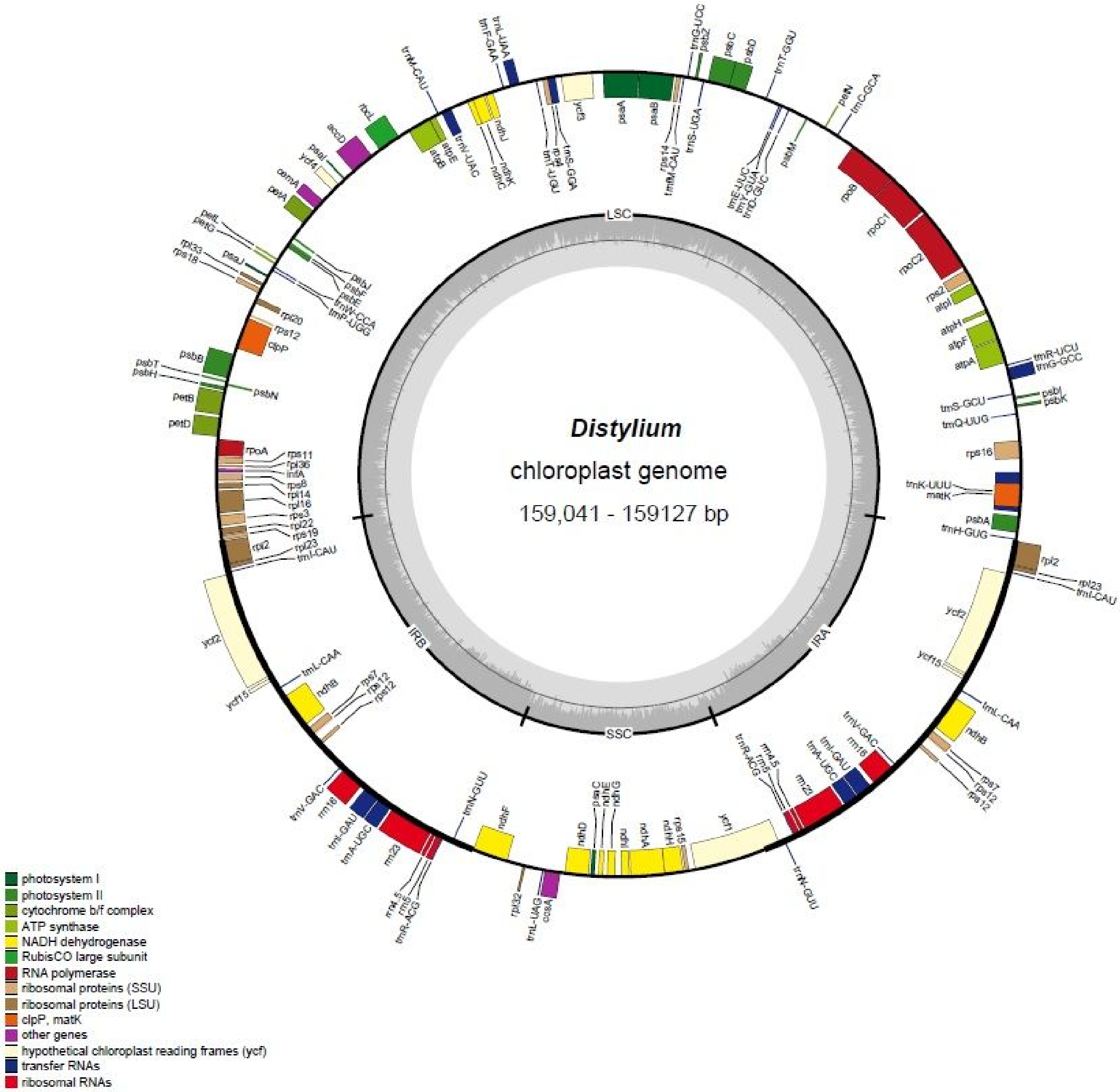

Figure 1

Gene map of the Distylium plastomes. Genes shown inside the inner circle are transcribed counterclockwise and those outside the circle are transcribed clockwise. The GC content of the genome is indicated by the dashed area in the inner circle.

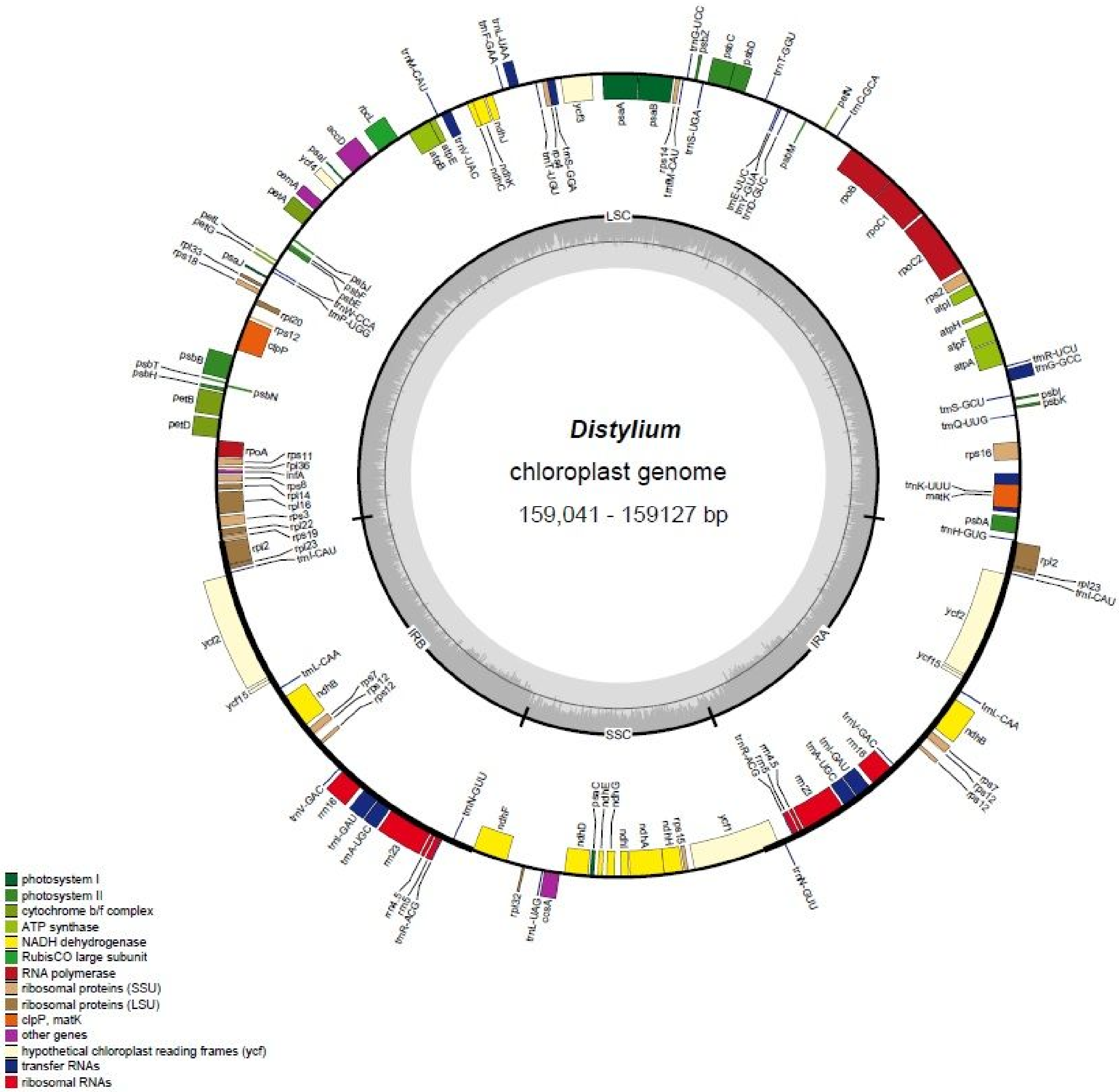

Figure 1

Gene map of the Distylium plastomes. Genes shown inside the inner circle are transcribed counterclockwise and those outside the circle are transcribed clockwise. The GC content of the genome is indicated by the dashed area in the inner circle.

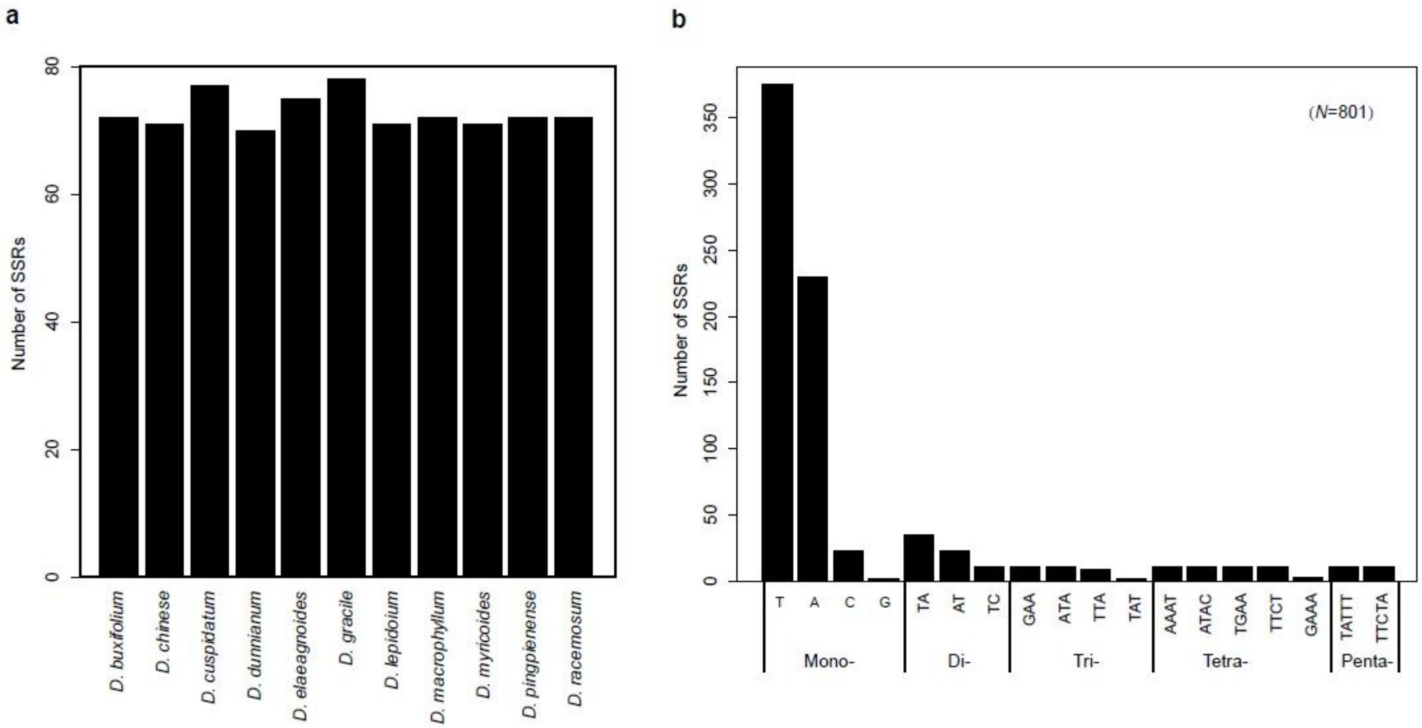

Figure 2

Frequency of the simple sequence repeat (SSR) sequences in the Distylium plastomes. a. The number of SSRs detected in the 11 Distylium species; b. Frequency of SSRs with di- to penta-nucleotide motifs.

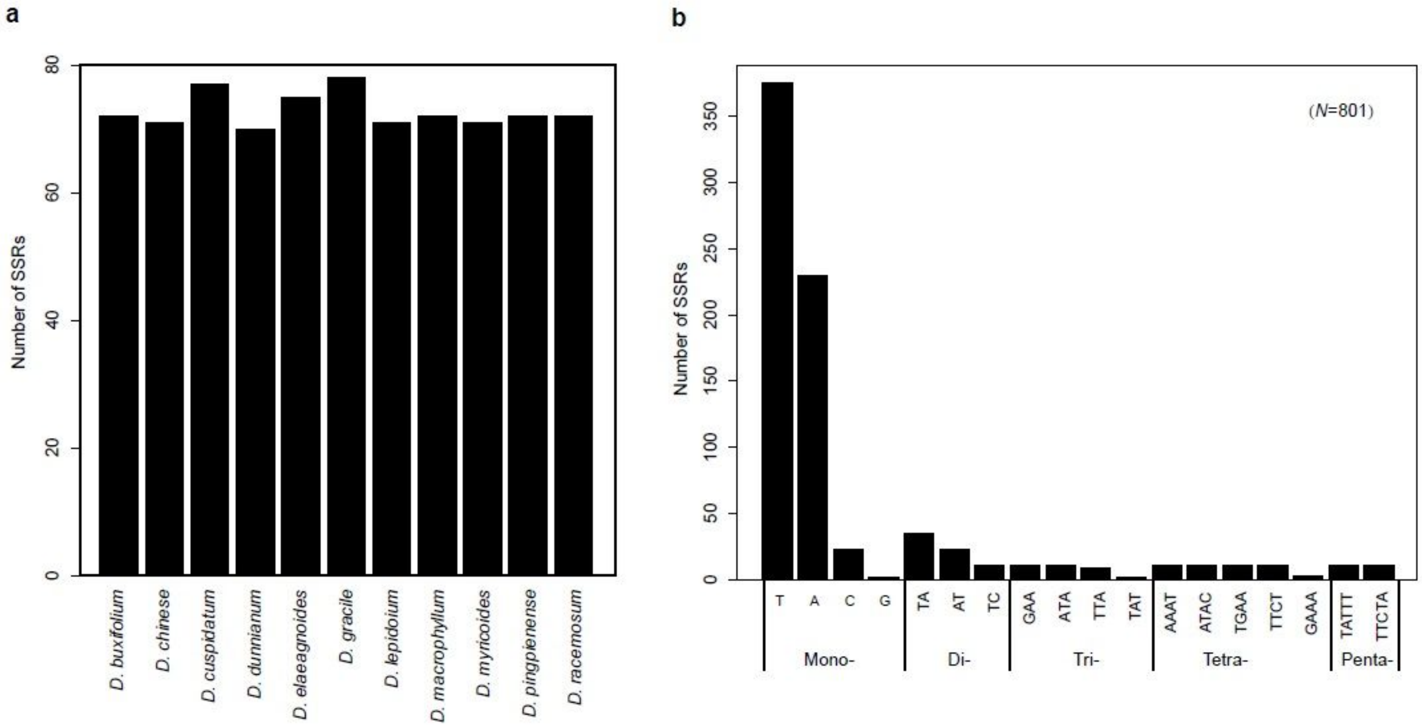

Figure 2

Frequency of the simple sequence repeat (SSR) sequences in the Distylium plastomes. a. The number of SSRs detected in the 11 Distylium species; b. Frequency of SSRs with di- to penta-nucleotide motifs.

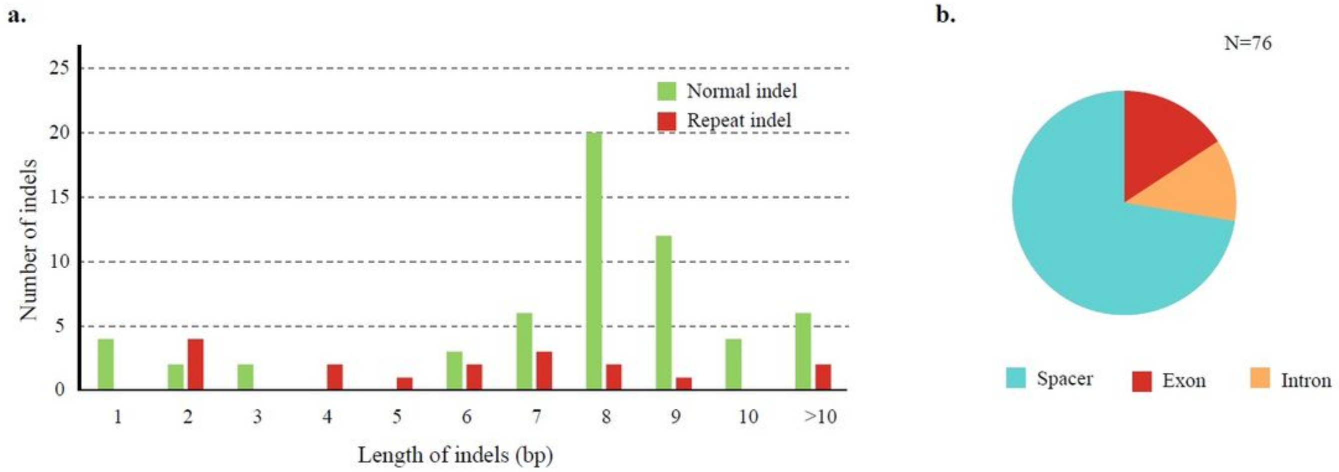

Figure 3

Analyses of indels in the Distylium plastomes. a. Number and size of the indels among the Distylium plastomes. b. Frequency of indel types and locations.

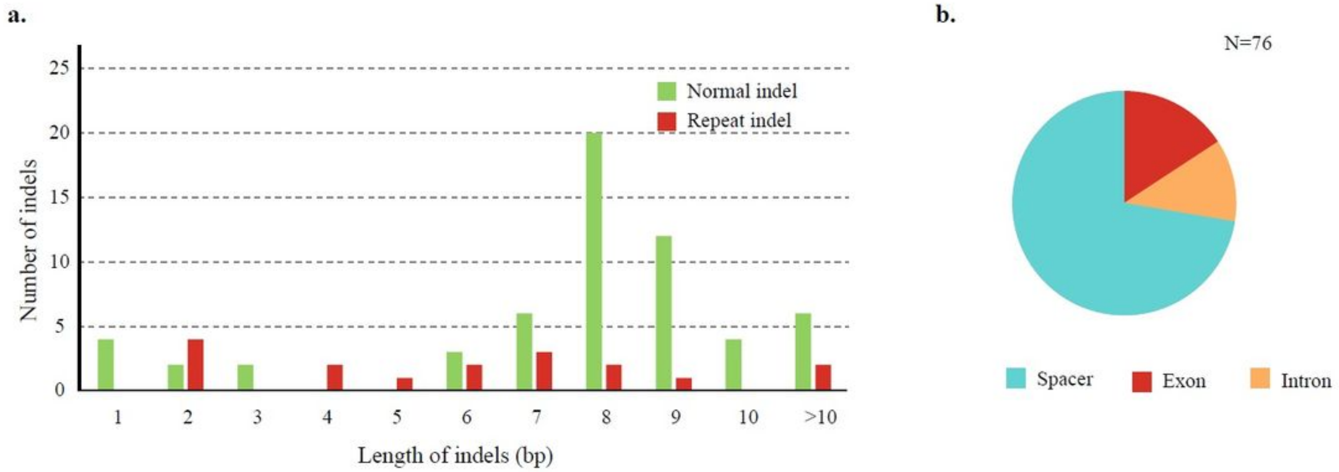

Figure 3

Analyses of indels in the Distylium plastomes. a. Number and size of the indels among the Distylium plastomes. b. Frequency of indel types and locations.

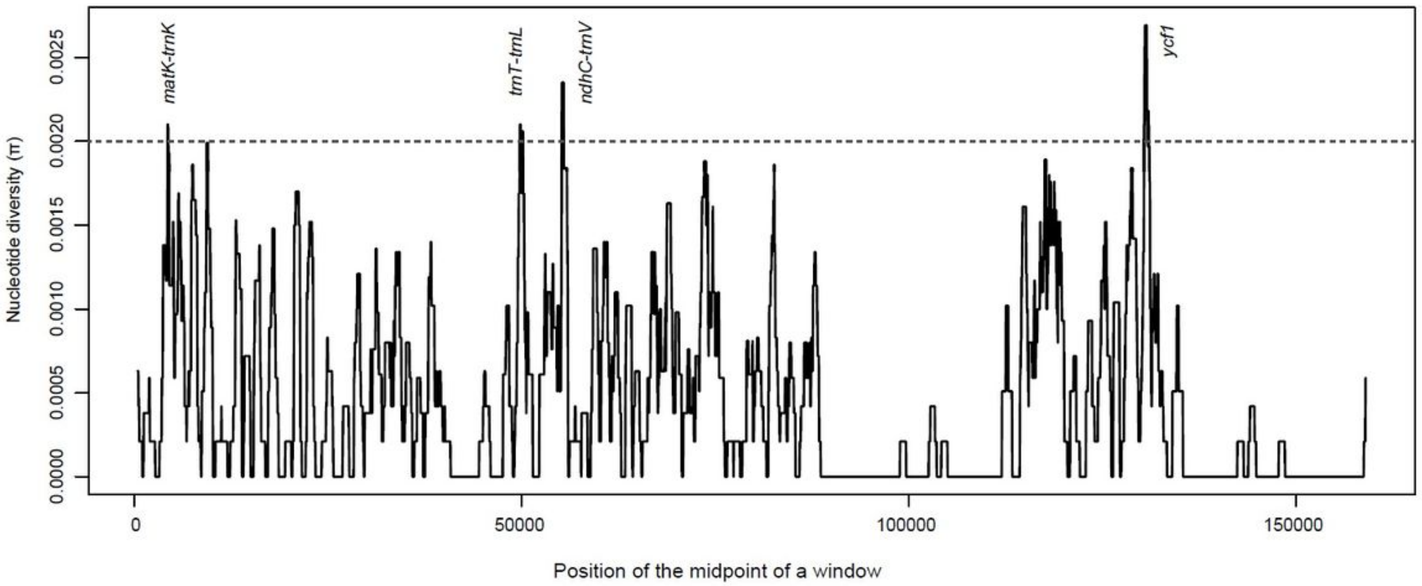

Figure 4

Nucleotide diversity (π) in the Distylium plastomes using sliding window method. The four mutation hotspot regions (π > 0.002) were annotated. π values were calculated in 800 bp sliding windows with 50 bp steps.

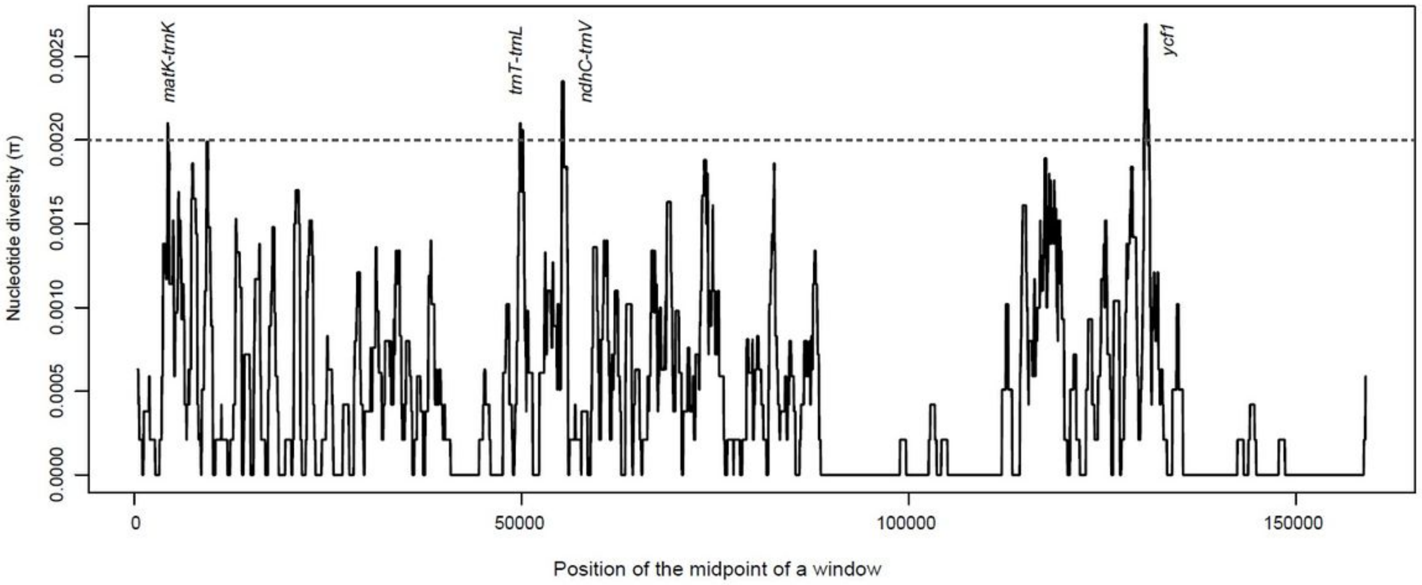

Figure 4

Nucleotide diversity (π) in the Distylium plastomes using sliding window method. The four mutation hotspot regions (π > 0.002) were annotated. π values were calculated in 800 bp sliding windows with 50 bp steps.

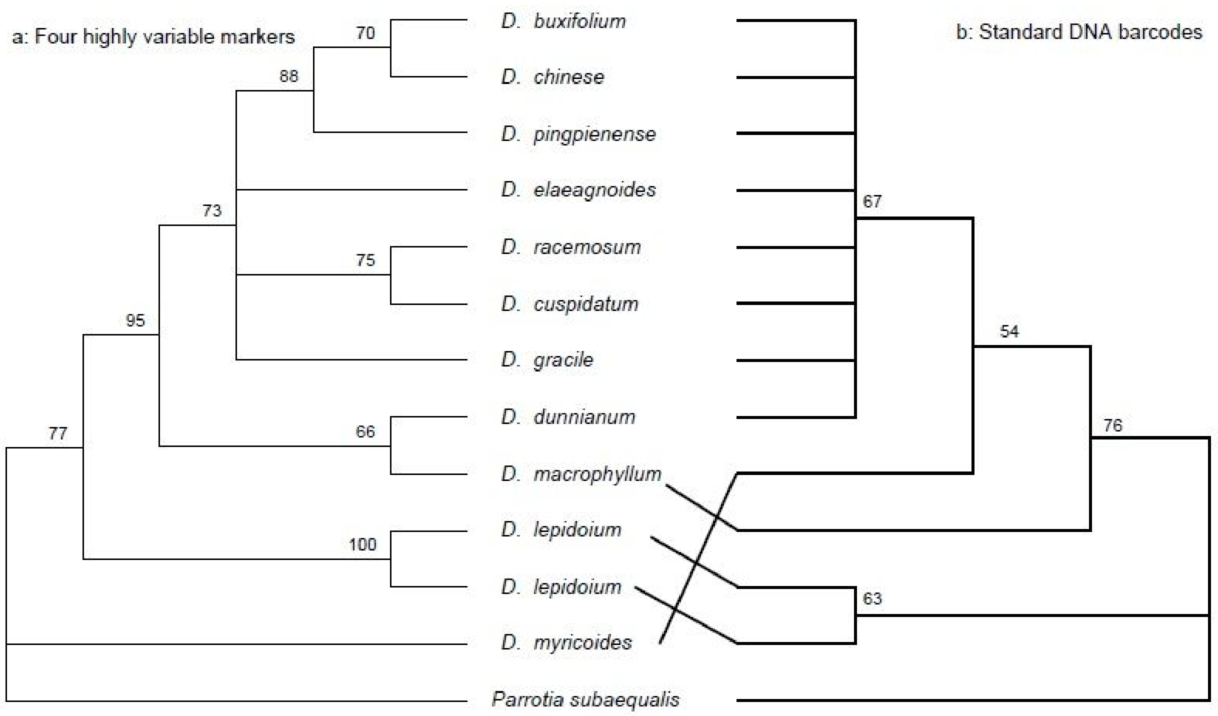

Figure 5

Neighbor joining tree for Distylium using combine three universal plant DNA barcodes and four highly variable regions combinations.

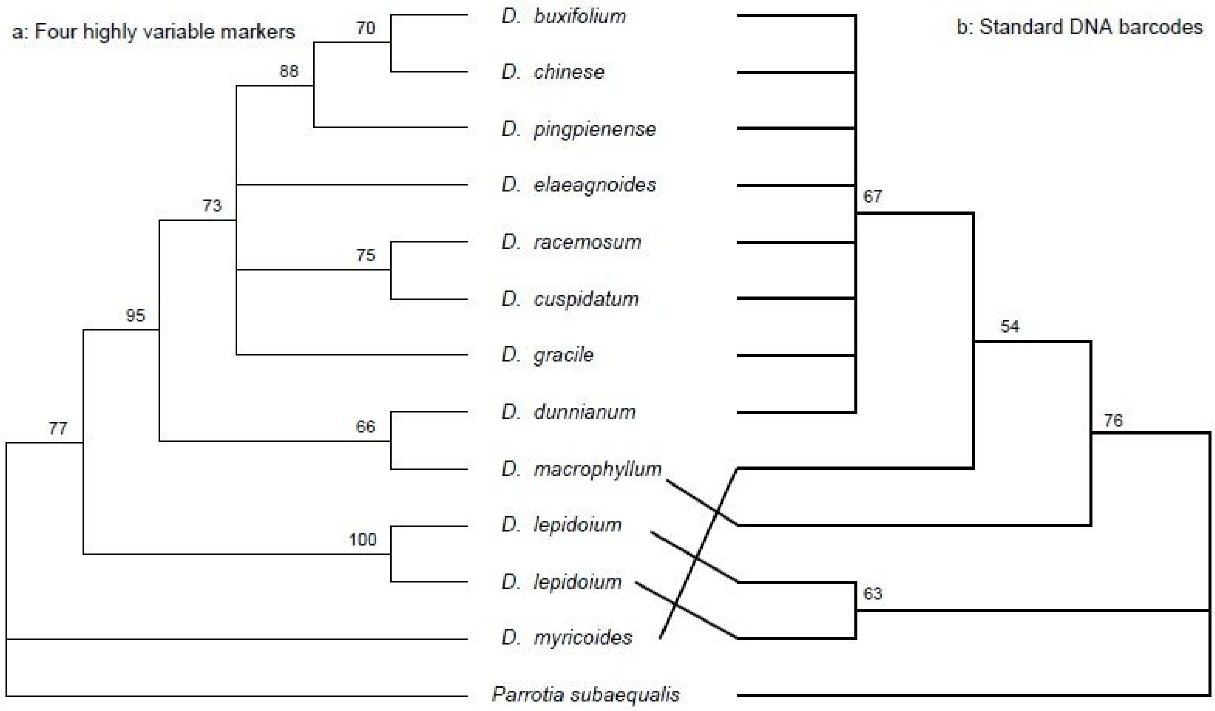

Figure 5

Neighbor joining tree for Distylium using combine three universal plant DNA barcodes and four highly variable regions combinations.

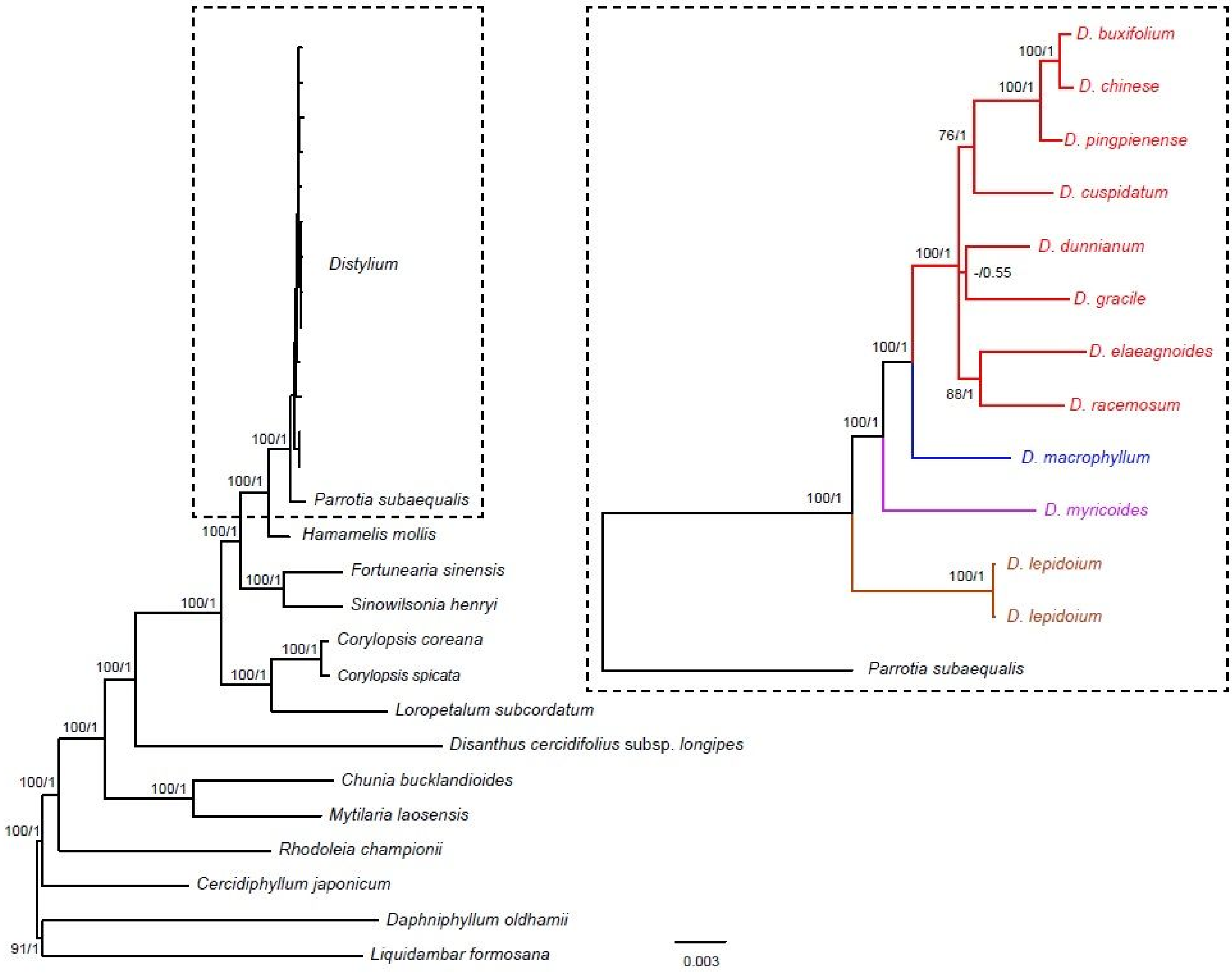

Figure 6

Phylogenetic reconstruction of Hamamelidaceae from maximum likelihood (ML) and Bayesian inference (BI) methods based on the plastome dataset. The ml tree is shown. Number of the branches represent ML bootstrap support value (BP) /Bayesian posterior probability (PP).

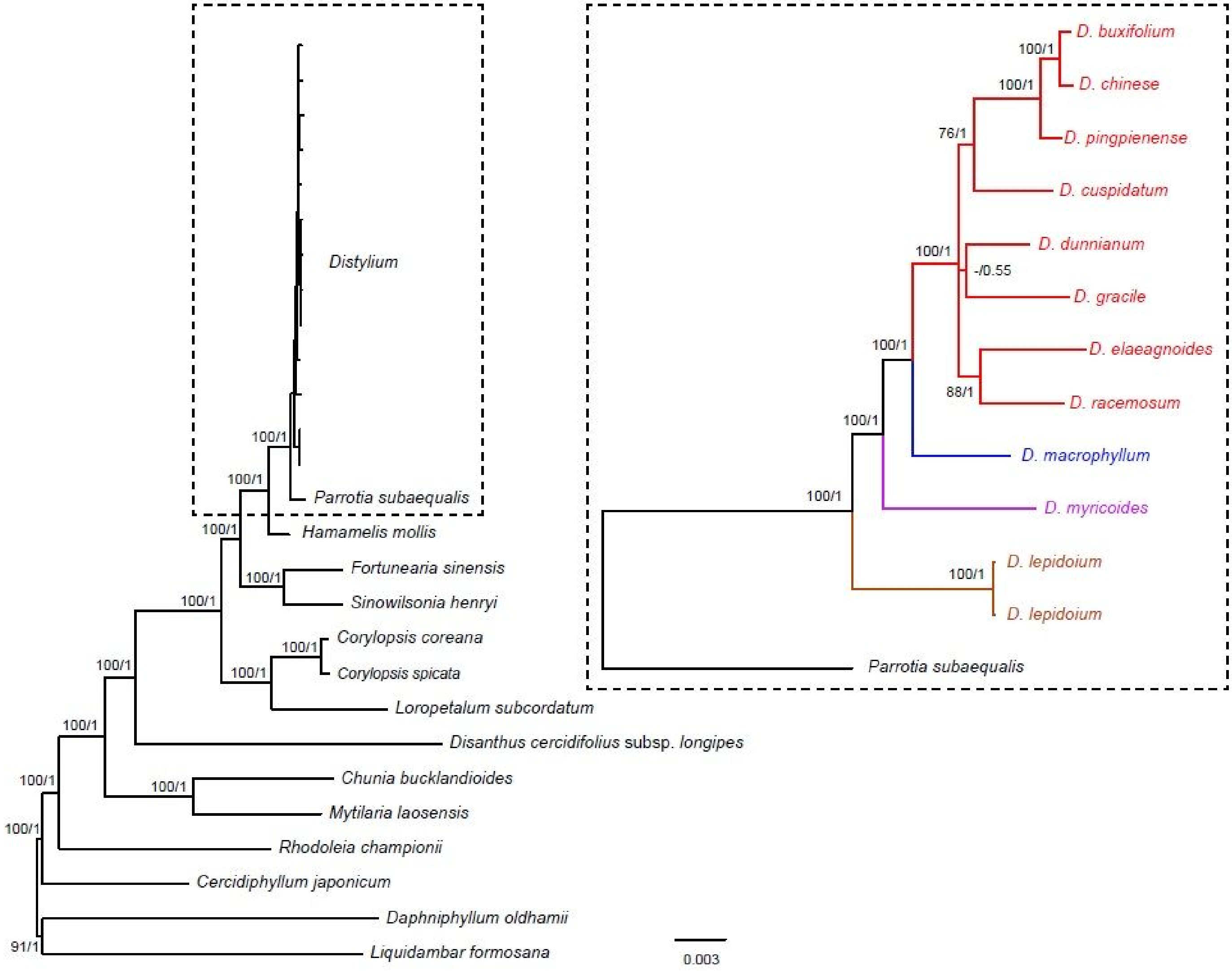

Figure 6

Phylogenetic reconstruction of Hamamelidaceae from maximum likelihood (ML) and Bayesian inference (BI) methods based on the plastome dataset. The ml tree is shown. Number of the branches represent ML bootstrap support value (BP) /Bayesian posterior probability (PP).

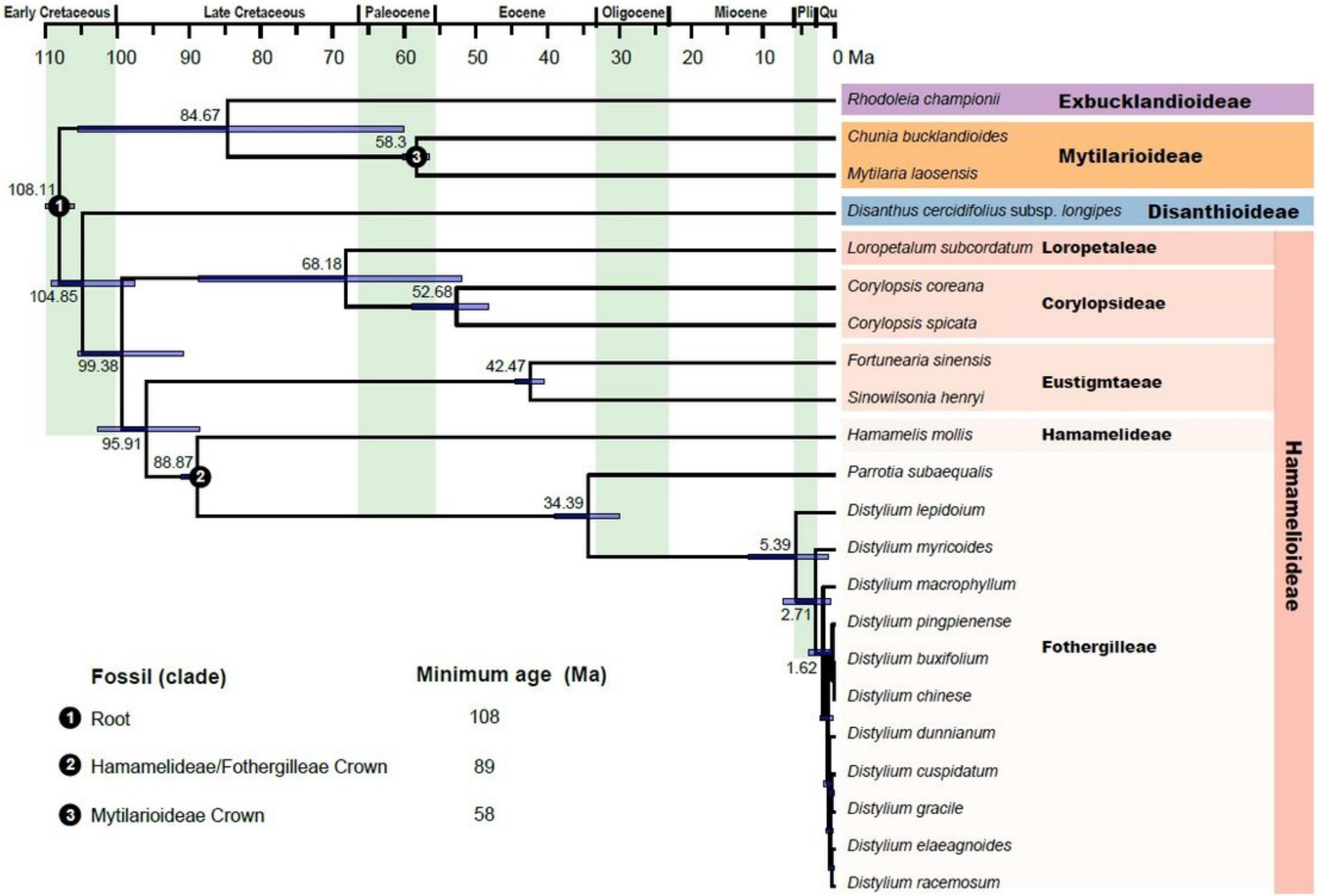

Figure 7

Divergence times of Hamamelidaceae obtained from BEAST analysis based on the complete plastome sequences. Mean divergence time of the nodes were shown next to the nodes while the blue bars correspond to the 95% highest posterior density (HPD). Black circles indicate the three calibration points.

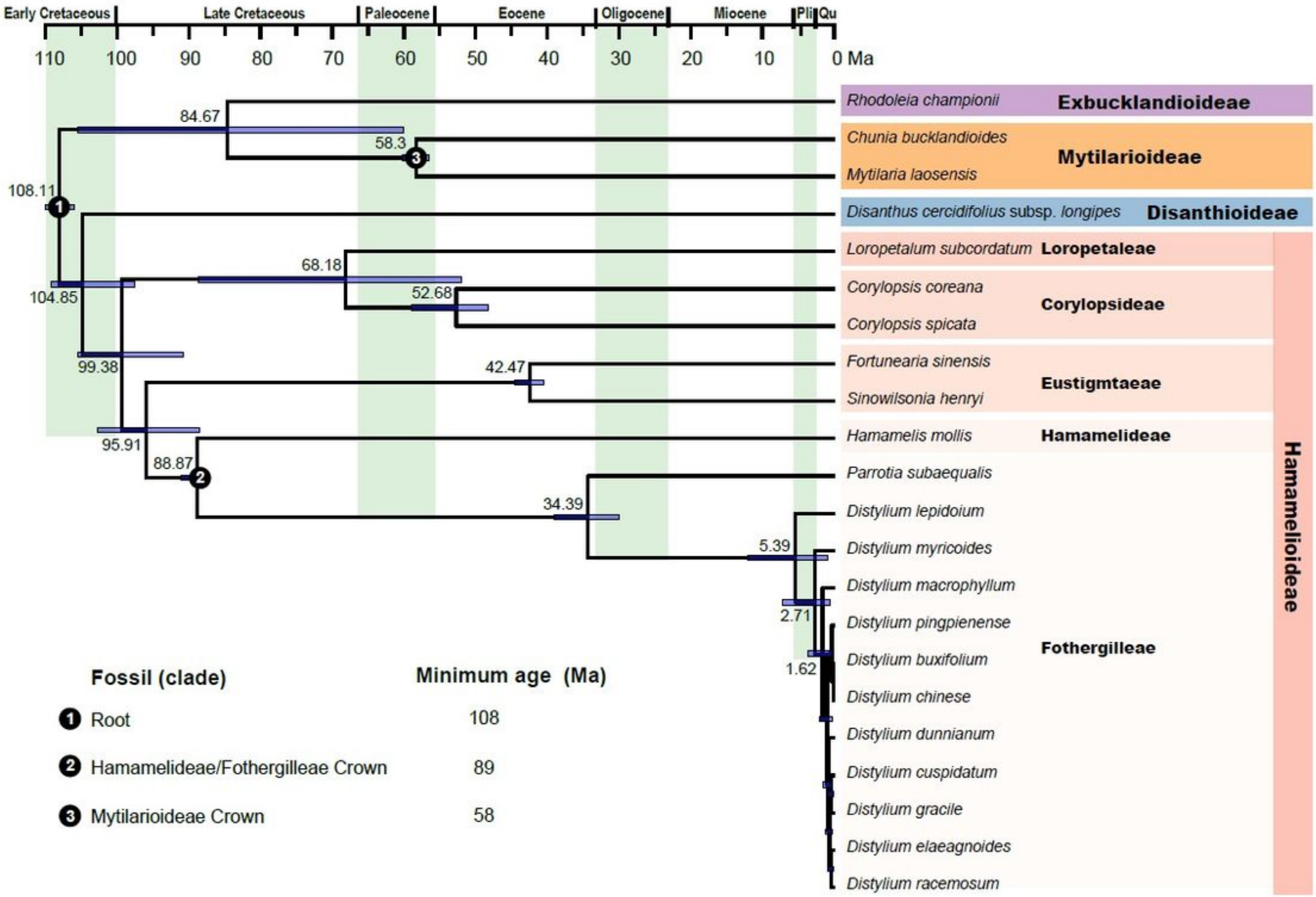

Figure 7

Divergence times of Hamamelidaceae obtained from BEAST analysis based on the complete plastome sequences. Mean divergence time of the nodes were shown next to the nodes while the blue bars correspond to the 95% highest posterior density (HPD). Black circles indicate the three calibration points.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS1samples.xlsx
- TableS1samples.xlsx
- TableS2.xlsx
- TableS2.xlsx