RESEARCH ARTICLE

Structure and features of the complete chloroplast genome of *Melastoma dodecandrum*

Xiasheng Zheng¹ · Changwei Ren² · Song Huang¹ · Jing Li³ · Ying Zhao³

Received: 26 September 2018 / Revised: 27 December 2018 / Accepted: 19 February 2019 / Published online: 12 March 2019
© Prof. H.S. Srivastava Foundation for Science and Society 2019

Abstract *Melastoma dodecandrum*, the only creeping species in the *Melastoma* genus, serves as a medicinal herb in southeast China. It belongs to the huge family Melastomataceae, which contains over 5000 species worldwide. In this study, we used next-generation sequencing to determine the complete chloroplast genome sequences of *M. dodecandrum*, which is a circular molecule of 156,611 bp in length. After annotation, we identified 131 putative genes in total, comprised of 85 protein-coding genes, 38 transfer RNA genes and 8 ribosomal RNA genes. Genome structure, GC content, repeat sequences and codon usage were investigated to gain a comprehensive understanding of this genome. Furthermore, we conducted comparative genome analyses between the *M.*

dodecandrum genome and that of four other Melastomataceae species. Additionally, a phylogenetic analysis was performed based on available chloroplast genomes of Melastomataceae species and several Myrtaceae species, revealing the taxonomic relationships between *M. dodecandrum* and related species. In conclusion, our study represents the first look into the complete chloroplast genome of *M. dodecandrum*, providing abundant information for further studies such as species identification, taxonomy and phylogenetic resolution of Melastomataceae species.

Keywords Chloroplast genome · *Melastoma dodecandrum* · Melastomataceae · Phylogenetic analysis · Next-generation sequencing

Xiasheng Zheng and Changwei Ren have contributed equally to this work.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s12298-019-00651-x>) contains supplementary material, which is available to authorized users.

✉ Jing Li
lijing@gzucm.edu.cn

✉ Ying Zhao
drzhaoying@126.com

¹ DNA Barcoding Laboratory for TCM Authentication, Mathematical Engineering Academy of Chinese Medicine, Guangzhou University of Chinese Medicine, Guangzhou 510006, China

² Department of Cardiovascular Surgery Center, Beijing Anzhen Hospital, Capital Medical University, Beijing 100029, China

³ Traditional Chinese Medicine Gynecology Laboratory in Lingnan Medical Research Center, Guangzhou University of Chinese Medicine, Guangzhou 510006, China

Abbreviations

cp	Chloroplast
gDNA	Genomic DNA
IR	Inverted repeat
LSC	Large single copy
ML	Maximum likelihood
NCBI	National Center for Biotechnology Information
NGS	Next generation sequencing
NJ	Neighbor-joining
SSC	Small single copy
SSR	Simple sequence repeats

Introduction

Melastoma dodecandrum Lour. is a creeping shrub widely distributed in southeast China, which prefers to grow in acidic soils (Sciences 1984). It is used in the clinic as an alternative medicine to cure several diseases in south

China, due to its multiple pharmaceutical properties, including hemostatic, anti-inflammatory (Ishii et al. 1999; Yang et al. 2014) and antibacterial (Liu et al. 2014). Compared to other species that are erect shrubs, *M. dodecandrum* is the only creeping species within the *Melastoma* genus. In addition, it was reported that *M. dodecandrum* could hybridize with three other congeneric species, indicating that strong reproductive isolation exists between *M. dodecandrum* and these species (Zou et al. 2017). Previously, microsatellite markers were designed and evaluated to study the genetic diversity and population structure of *M. dodecandrum* and its congeneric species (Liu et al. 2013; Renner and Meyer 2001; Zou et al. 2017). However, phylogenetic analysis based on one or several plastid markers has proven to be unsatisfactory, especially for the analysis of lineages within the Melastomataceae, which contains over 5000 species worldwide (Reginato et al. 2016). In contrast, phylogeny based on the complete chloroplast genome represents a significant improvement over phylogenies based on a single or a few plastid loci (Reginato et al. 2016).

The chloroplast (cp) is the characteristic organelle of a plant cell, which is believed to have descended from an ancestral cyanobacterium via endosymbiosis (Dyall et al. 2004). The chloroplast genome contains genetic information about the enzymic machinery necessary for photosynthesis (Green 2011) as well as the gene expression system of the chloroplast (Allen 2015). Moreover, cp genomes are generally small in size and conserved in structure and gene content (Wicke et al. 2011). For this reason, cp genomes have been increasingly used for multiple molecular studies on organelle biology (Jarvis and Lopez-Juez 2013; Smith and Keeling 2015), synthetic biology (Scharff and Bock 2014), development of molecular markers (Wu et al. 2010), phylogenetic analysis (Nikiforova et al. 2013) and so on. Due to increasing sequencing depth and decreasing costs, next-generation sequencing (NGS), has proven to be a powerful tool for the acquisition of new and complete cp genomes (Daniell et al. 2016; Moore et al. 2006). To date, there are more than 2000 cp genomes available in the National Center for Biotechnology Information (NCBI).

The objective of this study was to gain a comprehensive understanding of the cp genome of *M. dodecandrum* by describing its genome structure and features, and by comparing it to that of several congeneric species.

Results and discussion

Genome assembly

Using the Illumina HiSeq 2000 platform, a total of 25,225,380 paired-end raw reads were generated from a

sequencing library of genomic DNA of *M. dodecandrum*, accounting for 7.57 GB data in total. After data filtering by trimming adaptors and removing low-quality reads, 24,967,763 clean reads were retained for the following analysis. With the cp-DNA sequence of a closely related species, *M. candidum*, serving as a reference, cp-like reads were extracted from the clean reads and then assembled into the complete cp genome sequence of *M. dodecandrum* (designated as Md_cp genome). The Md_cp genome was annotated and deposited into GenBank with the Accession Number MH748092.

Analysis of genome features

The Md_cp genome is a circular molecule of 156,611 bp with the classical quadripartite organization commonly observed for plant chloroplasts, which contains one large single copy (LSC), one small single copy (SSC), as well as two inverted repeat (IR) regions. A circular map of the Md_cp genome is presented in Fig. 1. The GC content of the Md_cp genome is 37.1 mol%, which is consistent with other members of Melastomataceae (Ng et al. 2018; Reginato et al. 2016). A further analysis of the four regions of the Md_cp genome (see Table 1) revealed that the GC content of the IR regions is greater than that of the two single copy regions. In addition, we observed a bias towards a high AT representation at the third codon position, similar to what has been found in other land plant cp genomes (Shen et al. 2017), which is believed to be beneficial for distinguishing cp-DNA from nuclear and mitochondrial DNA (Clegg et al. 1994).

After annotation, we identified a total of 131 predicted genes (including duplicates and *ycf* genes) in the Md_cp genome, comprised of 114 unique genes and 17 duplicates in the IR regions. The unique genes included 79 protein-coding genes (CDS), 31 transfer RNAs (tRNA) and 4 ribosomal RNAs (rRNA), while 6 CDS, 7 tRNA and 4 rRNA gene duplicates were found in the IR regions (see Table 2). In general, the Md_cp genome shares similarity to that of its sister species *M. candidum* in terms of gene content and gene order (Ng et al. 2018). Among the predicted genes, nine CDS (*atpF*, *ndhA*, *ndhB*, *petB*, *petD*, *rpl16*, *rpl2*, *rpoC1* and *rps16*) and five tRNAs contain one intron, while three CDS (*clpP*, *rps12* and *ycf3*) contain two introns (see Table 3). Based on the annotation of the reference *M. candidum* cp genome, homologous sequences of *ycf15* were found in the Md_cp genome. Nonetheless, those sequences share poor homology to other reported sequences (Shi et al. 2013). Thus, we eliminated those putative *ycf15* annotations in the Md_cp genome.

Fig. 1 Map of the *Melastoma dodecandrum* chloroplast genome. Genes in different functional groups are color coded as shown in the legend (bottom left). Genes shown outside the outer circle are

transcribed counterclockwise, while those inside are transcribed clockwise. The shaded area inside the inner circle indicates the GC content, with dark shading indicating a higher GC mol percent

Analysis of codon usage and RNA editing sites

Based on the sequences of CDS, codon usage frequency of the Md_cp genome was evaluated and summarized (see Supporting Information Table S1). In general, all 61 codons, as well as three stop codons, were found in the Md_cp genome, encoding 20 amino acids. In sum, the CDS of the Md_cp genome is encoded by 26,770 codons. Among them, 2885 and 311 codons encode leucine and

cysteine, respectively, representing the most and the least frequently used amino acids. Relative synonymous codon usage (RSCU), whose value reflects the ratio between the usage frequency and expected frequency of a particular codon, was introduced to evaluate the codon usage of the Md_cp genome. As illustrated in Fig. 2, the RSCU value of the Md_cp genome exhibits a pattern similar to that seen in four other Melastomataceae species. We found a strong AT bias in codon usage, with an RSCU value > 1, including

Table 1 Base composition in the *Melastoma dodecandrum* chloroplast genome

Region	T/U (%)	C (%)	A (%)	G (%)	Total (bp)
LSC	33.2	17.9	31.8	17.2	86,015
IRA	28.5	20.5	29.0	22.1	26,750
SSC	34.0	16.2	34.8	15.0	17,096
IRB	28.9	22.1	28.5	20.5	26,750
Total	31.7	18.8	31.1	18.3	156,611
CDS	31.5	17.6	30.8	20.0	80,310
1st position	24.1	18.9	30.8	26.2	26,770
2nd position	32.8	19.8	30.0	17.5	26,770
3rd position	37.7	14.3	31.6	16.4	26,770

UUA (1.90) in leucine, the stop-codon UAA (1.80) and UAU (1.64) in tyrosine, which is a common phenomenon in land plant plastomes (Morton 1998).

Additionally, a total of 48 putative RNA editing sites were identified and summarized (see Supporting Information Table S2), based on 35 CDS sequences. The most frequent occurrence (15) was the amino acid conversion from leucine (L) to serine (S). Meanwhile, the conversion from methionine (M) to threonine (T) was observed twice, in *ndhD* and *rpl2*. It has been reported that the ACG codon, with which *ndhD* starts, may be restored to a canonical start codon AUG by RNA editing (Hoch et al. 1991). Meanwhile, we found that *rps19* starts with GUG,

consistent with what has been previously observed in date palm and 14 other higher plants (Yang et al. 2010). Furthermore, the non-canonical start codon GUG was also found in *psbC* of the tobacco plastid, suggesting that a strong interaction between the extended Shine-Dalgarno-like sequence (GAGGAGGU, nine nucleotides upstream of the GUG) and the 3' end of 16S rRNA facilitates translation initiation from the GUG (Kuroda et al. 2007).

Analysis of sequence repeats

Sequence repeats, including simple sequence repeats (SSRs) and long repeats, of the Md_cp genome were also investigated. Simple sequence repeats, also known as DNA microsatellites, refer to short (1–6 bp) tandemly repeated DNA sequences, which are widely distributed throughout nuclear genome. SSRs have been utilized for multiple genetic analyses, including plant population genetics, as well as evolutionary and ecological studies (Huang et al. 2014; Zhao et al. 2015). We found 308 SSRs found in the Md_cp genome in total (see Table 4), among which mononucleotide SSRs account for 53.6%. It is worth noting that these SSRs are generally composed of tandem adenine (A) or thymine (T), which contributes to the AT richness of the cp genome and is consistent with other species within the Melastomataceae family.

In addition to the SSRs, we also examined long repeats of the Md_cp genome, including forward, reverse, complement and palindromic repeats, with a minimum detected

Table 2 Gene contents in the chloroplast genome of *Melastoma dodecandrum*

Classification of genes	Gene names	Number of genes
Photosystem I	<i>psaA, psaB, psaC, psaI, psaJ</i>	5
Photosystem II	<i>psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ</i>	15
Cytochrome b/f complex	<i>petA, petB, petD, petG, petL, petN</i>	6
ATP synthase	<i>atpA, atpB, atpE, atpF, atpH, atpI</i>	6
NADH dehydrogenase	<i>ndhA, ndhB</i> (× 2), <i>ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK</i>	12
RubisCO large subunit	<i>rbcL</i>	1
RNA polymerase	<i>rpoA, rpoB, rpoC1, rpoC2</i>	4
Ribosomal proteins (in SSC region)	<i>rps2, rps3, rps4, rps7</i> (× 2), <i>rps8, rps11, rps12</i> (× 2), <i>rps14, rps15, rps16, rps18, rps19</i>	14
Ribosomal proteins (in LSC region)	<i>rpl2</i> (× 2), <i>rpl14, rpl16, rpl20, rpl22, rpl23</i> (× 2), <i>rpl32, rpl33, rpl36</i>	11
Ribosomal RNAs	<i>rrn 4.5</i> (× 2), <i>rrn 5</i> (× 2), <i>rrn 16</i> (× 2), <i>rrn 23</i> (× 2)	8
Protein of unknown function	<i>ycf1</i> (× 2), <i>ycf2</i> (× 2), <i>ycf3, ycf4</i>	6
Transfer RNAs	tRNAs (7 contain an intron, 7 in the inverted repeats region)	38
Other genes	<i>accD, ccsA, cemA, clpP, matK</i>	5
Total		131

Table 3 Genes with introns in the chloroplast genome of *Melastoma dodecandrum*

Gene	Location	Exon I (bp)	Intron I (bp)	Exon II (bp)	Intron II (bp)	Exon III (bp)
<i>atpF</i>	LSC	144	826	411		
<i>clpP</i>	LSC	71	951	292	632	228
<i>ndhA</i>	SSC	549	1015	543		
<i>ndhB</i>	IR	780	680	756		
<i>petB</i>	LSC	6	774	642		
<i>petD</i>	LSC	8	794	475		
<i>rpl16</i>	LSC	9	1039	399		
<i>rpl2</i>	IR	391	662	434		
<i>rpoC1</i>	LSC	430	726	1619		
<i>rps12</i>	IR	114	–	232	540	26
<i>rps16</i>	LSC	39	817	216		
<i>trnA-UGC</i>	IR	38	801	35		
<i>trnI-GAU</i>	IR	42	842	35		
<i>trnK-UUU</i>	LSC	37	2537	35		
<i>trnL-UAA</i>	LSC	37	538	50		
<i>trnV-UAC</i>	LSC	39	611	37		
<i>ycf3</i>	LSC	126	740	226	823	155

Fig. 2 Codon usage of 20 amino acid and stop codons based on all protein-coding genes in the chloroplast genome of *Melastoma dodecandrum* and four related species. There are five bars for every amino acid and the stop codons, representing the RSCU value of

chloroplast genome of *M. dodecandrum*, *M. candidum*, *Allomaieta villosa*, *Barthea barthei* and *Tigridiopalma magnifica* from left to right. Different codons that encode the same amino acid are colored differently

length set at 30 bp. The cp genomes of four other Melastomataceae species were added to the comparison in this analysis, and the results are shown in Fig. 3. In general,

there are 49 long repeats in both *M. dodecandrum* and *M. candidum*, while less than half of this number is found in *Allomaieta billos* and *Tigridiopalma magnifica*. Most of

Table 4 Types and amounts of SSRs in the chloroplast genomes of *Melastoma dodecandrum* and *M. candidum*

SSR type	Repeats unit	Amount		Ratio	
		<i>M. dodecandrum</i>	<i>M. candidum</i>	<i>M. dodecandrum</i> (%)	<i>M. candidum</i> (%)
Mono	A/T	163	165	98.8	98.8
	C/G	2	2	1.2	1.2
Di	AC/GT	1	1	1.6	1.6
	AG/CT	20	20	32.8	32.8
Tri	AT/AT	40	40	65.6	65.6
	AAC/GTT	8	8	11.3	10.8
	AAG/CTT	19	18	26.8	24.3
	AAT/ATT	29	33	40.8	44.6
	ACC/GGT	2	2	2.8	2.7
	ACT/AGT	3	3	4.2	4.1
	AGC/CTG	4	4	5.6	5.4
	AGG/CCT	1	1	1.4	1.4
	ATC/ATG	5	5	7.0	6.8
	AAAT/ATTT	7	5	77.8	71.4
	AATT/AATT	1	1	11.1	14.3
	AGAT/ATCT	1	1	11.1	14.3
Penta	AAAAT/ATTTT	1	0	50.0	0.0
	AATAT/ATATT	1	0	50.0	0.0

these long repeats are forward, palindromic, and 30–49 bp in length. These sequence repeats provide valuable information for further phylogenetic studies on Melastomataceae species.

Expansion and contraction of the inverted repeats regions

The IRs are the most conserved regions in cp genomes. Therefore, expansion and contraction of their borders, which are common evolutionary events, are often investigated to assess the structural differences among cp genomes (Kim and Lee 2004; Wang et al. 2008). Genes that span or are in close proximity to the neighboring regions of the cp genomes of *M. dodecandrum* and four other Melastomataceae species were compared and are illustrated in Fig. 4. The boundaries of LSC-IRB are located in the *rps19* genes, while those of IRB-SSC spanned the *ndhF* genes. As a result, two gene fragments, namely $\psi ndhF$ and $\psi rps19$, were also found to be located in the IRA region. Moreover, the 5' truncated *rps19* gene was located downstream of *rpl2* in the IRA region and was followed by an intact *trnH* in the LSC, indicating that the IRA-LSC junction in Melastomataceae is in agreement with those of other eudicots (Wang et al. 2008). At the same time, we observed that *ycf1*, which spans across the SSC-IRA junction, produces a corresponding pseudogene, 5'

truncated *ycf1*, in the IRB region. Judging from Fig. 4, a very similar pattern of expansion and contraction of the IR regions was observed among the Melastomataceae species, indicating that structural conservation is common within the cp genomes of Melastomataceae.

Genome comparison analysis

Comparative analysis of different genomes allows the examination of how DNA sequences diverge between and/or among related species. We performed comparative analysis of the cp genome of *M. dodecandrum* with four reported Melastomataceae species, including *M. candidum* (NC_034716), *Allomaieta villosa* (NC_031875), *Barthea barthei* (NC_035661) and *Tigridiopalma magnifica* (NC_036021), using the mVISTA program (see Fig. 5). Among them, the cp genome of *M. dodecandrum* represents the second largest, which is 71 bp shorter than that of *M. candidum* with a length of 156,682 bp. As shown in Fig. 5, there are slight sequence differences between the cp genomes of the two *Melastoma* species. Furthermore, we observed more sequence divergence in the LSC and SSC regions of the other three cp genomes, which mainly occur in the intergenic spacers. Moreover, divergence also occurs in several protein-coding genes, such as *clpP*, *matK*, *rpoC2*, *ycf1* and *ycf3*. The sequence differences identified in the comparison provide potential spots for biomarker

Fig. 3 Distribution of long repeat sequences in the chloroplast genome of *Melastoma dodecandrum* and four related species. *F* forward repeats, *P* palindrome repeats, *R* reverse repeats and

C complement repeats. Repeats with different lengths are colored differently as indicated (color figure online)

discovery and species authentication. In addition, both IR regions of the compared cp genomes, which contain four rRNAs and *ycf2*, appear to be very conserved in sequence.

Phylogenetic analysis

As mentioned above, cp genomes are useful for phylogenetic analyses in plants. In order to investigate the taxonomic relationship of *M. dodecandrum* and other related species, we performed a phylogenetic study based on the cp genome sequences of 20 Melastomataceae species, along with three Myrtaceae species as an outgroup, using both maximum-likelihood (ML) and neighbor-joining (NJ) trees. As shown in Fig. 6, both trees exhibit the same topological pattern. In detail, the two *Melastoma* species grouped together, and further clustered with the other genus of Melastomataceae. The Melastomataceae branch could then be clearly distinguished from the three outgroup species of Myrtaceae. Bootstrap support of most nodes in this phylogenetic tree is high. This phylogeny is consistent with that of a previous report (Reginato et al. 2016), which was reconstructed based on 78 protein-coding genes, and thus extends to a broader taxonomic range with two

Melastoma species and one *Tigridiopalma* species. Furthermore, the consistency of these two phylogeny results demonstrates that whole cp genome sequences (including one IR region) are useful for phylogenetic analyses. It may be more convenient to use whole cp genome sequences rather than just coding region sequences, which require additional data extraction. Melastomataceae has been historically treated as a major core family of the Myrtales in taxonomic order. However, the limited number of available whole cp genome sequences has hampered a comprehensive assessment of the evolutionary position of Myrtales plants.

Materials and methods

Plant material, DNA extraction and sequencing

Genomic DNA (gDNA) was extracted from fresh tender leaves of a *M. dodecandrum* plant grown in the Medicinal Botanical Garden of Guangzhou University of Chinese Medicine, using a DNeasy Plant Mini Kit (Qiagen, Germany). The purity and integrity of the extracted gDNA

Fig. 4 Distribution of genes located in/near the borders of two IR regions in the chloroplast genome of *Melastoma dodecandrum* and four related species. The number around the gene features indicates the distance from the borders site to the other pole. Ψ stands for pseudogene

were measured by ultraviolet spectrophotometry and gel electrophoresis, respectively. DNA samples of good integrity and with both OD260/280 and OD260/230 ratios greater than 1.8 were used in subsequent experiments. After being sheared into 250 bp fragments, a sequencing library was constructed with this gDNA and submitted for NGS on an Illumina HiSeq 2000 platform to generate paired-end reads.

Genome assembly and annotation

Approximately 7.57 GB of resulting sequencing data was filtered and trimmed to clean reads with an average length of 130 bp. Then, the complete sequence of the *M. candidum* chloroplast genome was used as a reference, to extract cp-like reads from the clean reads of *M. dodecandrum*. Next, those cp-like reads were assembled using the Abyss 2.0 program (Jackman et al. 2016) to form a complete chloroplast genome sequence of *M. dodecandrum*. Self-alignment was performed using BLASTn to locate the precise position of the quadripartite structure. To verify the assembly, four junction regions between the IR regions and the LSC/SSC region were confirmed by PCR amplification. Annotation of the cp genome was performed using the GeSeq online tool (Tillich et al. 2017) (<https://chlorobox.>

[mpimp-golm.mpg.de/geseq.html](https://chlorobox.mpiwg-berlin.mpg.de/geseq.html)) with default parameters, which annotates not only protein-coding genes, but also tRNAs and rRNAs. Further examination and revision of the annotation information were manually performed under the assistance of the CLC Sequence Viewer (version 8). A circular map of the *M. dodecandrum* cp genome was drawn using the OGDRAW program (Lohse et al. 2013) (<https://chlorobox.mpiwg-berlin.mpg.de/OGDraw.html>).

Genome features analysis and genome comparison

After annotation, the Md_cp genome was analyzed for structural characteristics, including GC content and codon usage, using MEGA (version 7) (Kumar et al. 2016). Analysis of SSRs and long sequence repeats analysis was carried out using MISA (Hennequin et al. 2001) and REPuter (Kurtz et al. 2001) programs, respectively. Another online tool, the Predictive RNA Editor for Plants (PREP) suite (Mower 2009), was used to identify potential RNA editing sites, with a cutoff value set at 0.8. The mVISTA program (Frazer et al. 2004) was used for whole-genome alignment of the cp genomes of *M. dodecandrum* and four other Melastomataceae species.

Fig. 5 Genome comparison of *Melastoma dodecandrum* and four Melastomataceae species using the mVISTA program. Gray arrows indicate the position and orientation of genes; purple bars represent

exons; pink bars represent non-coding sequences. Y axis scales represent percentage identity ranging from 50 to 100% (color figure online)

Phylogenetic analysis

Phylogenetic analysis was performed using the whole cp genome sequences of *M. dodecandrum*, along with other 19 complete cp genome sequences within the same family and three of Myrtaceae as an outgroup, obtained from the NCBI Organelle Genome Resources database (see Supporting Information Table S3). Those cp genomes were trimmed manually to remove the IRA region, and then aligned with the MAFFT (version 7) (Katoh et al. 2002) program using the FFT-NS-i $\times$ 1000 strategy. The alignment information was then used to reconstruct both maximum-likelihood and neighbor-joining trees using MEGA, employing the GTR+G model and the Kimura 2-parameter model, respectively. Support was estimated through 1000 bootstrap replicates to assess the reliability of the phylogenetic tree.

Conclusions

The chloroplast genome serves as a useful tool for plant phylogenetic studies due to high conservation of gene content and order, and the virtual lack of recombination (Ravi et al. 2008). Here, we provide the complete chloroplast genome sequence of *M. dodecandrum*, which is a circular molecule of 156,611 bp with a classic quadripartite structure. Compared to other species within the family, the cp genome of *M. dodecandrum* exhibits high structural conservation, in terms of gene content, gene order and other structural features, which is consistent with the results of other reported Melastomataceae species (Reginato et al. 2016), and even matches that of the Myrtales (Gu et al. 2016). Importantly, the full sequence and annotation of the *M. dodecandrum* chloroplast genome will accelerate phylogenetic, population and

Fig. 6 Phylogenetic trees of *Melastoma dodecandrum* and 22 related species based on their chloroplast genome sequences. Maximum-likelihood tree (a); neighbor-joining tree (b). Bootstrap support value are given adjacent to the nodes, representing the percentages of resampling. Filled circle indicates the target species of this study; open circle indicates outgroup species. The bar (bottom left) indicates 0.01 changes per nucleotide position

evolutionary studies of this species, and further contribute to a better understanding of plastid biology among Melastomataceae.

References

- Allen JF (2015) Why chloroplasts and mitochondria retain their own genomes and genetic systems: colocation for redox regulation of gene expression. *Proc Natl Acad Sci USA* 112:10231–10238

- Clegg MT, Gaut BS, Learn GH Jr, Morton BR (1994) Rates and patterns of chloroplast DNA evolution. *Proc Natl Acad Sci USA* 91:6795–6801
- Daniell H, Lin CS, Yu M, Chang WJ (2016) Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biol* 17:134. <https://doi.org/10.1186/s13059-016-1004-2>
- Dyall SD, Brown MT, Johnson PJ (2004) Ancient invasions: from endosymbionts to organelles. *Science* 304:253–257. <https://doi.org/10.1126/science.1094884>
- Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I (2004) VISTA: computational tools for comparative genomics. *Nucleic Acids Res* 32:273–279. <https://doi.org/10.1093/nar/gkh458>
- Green BR (2011) Chloroplast genomes of photosynthetic eukaryotes. *Plant J* 66:34–44. <https://doi.org/10.1111/j.1365-3113X.2011.04541.x>
- Gu C, Tembrock LR, Johnson NG, Simmons MP, Wu Z (2016) The Complete Plastid Genome of *Lagerstroemia fauriei* and Loss of *rpl2* Intron from *Lagerstroemia* (Lythraceae). *PLoS ONE* 11:e0150752. <https://doi.org/10.1371/journal.pone.0150752>
- Hennequin C, Thierry A, Richard GF, Lecointre G, Nguyen HV, Gaillardin C, Dujon B (2001) Microsatellite typing as a new tool for identification of *Saccharomyces cerevisiae* strains. *J Clin Microbiol* 39:551–559. <https://doi.org/10.1128/JCM.39.2.551-559.2001>
- Hoch B, Maier RM, Appel K, Igloi GL, Kossel H (1991) Editing of a chloroplast mRNA by creation of an initiation codon. *Nature* 353:178–180. <https://doi.org/10.1038/353178a0>
- Huang H, Shi C, Liu Y, Mao SY, Gao LZ (2014) Thirteen *Camellia* chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. *BMC Evol Biol* 14:151. <https://doi.org/10.1186/1471-2148-14-151>
- Ishii R, Saito K, Horie M, Shibano T, Kitanaka S, Amano F (1999) Inhibitory effects of hydrolyzable tannins from *Melastoma dodecandrum* Lour. on nitric oxide production by a murine macrophage-like cell line, RAW264.7, activated with lipopolysaccharide and interferon-gamma. *Biol Pharm Bull* 22:647–653
- Jackman SD et al (2016) ABySS 20: resource-efficient assembly of large genomes using a Bloom filter. *BioRxiv* 27:768–777. <https://doi.org/10.1101/068338>
- Jarvis P, Lopez-Juez E (2013) Biogenesis and homeostasis of chloroplasts and other plastids. *Nat Rev Mol Cell Biol* 14:787–802. <https://doi.org/10.1038/nrm3702>
- Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30:3059–3066
- Kim KJ, Lee HL (2004) Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. *DNA Res* 11:247–261
- Kumar S, Stecher G, Tamura K (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol* 33:1870–1874. <https://doi.org/10.1093/molbev/msw054>
- Kuroda H, Suzuki H, Kusumegi T, Kusumegi T, Hirose T, Hirose T, Yukawa Y, Yukawa Y, Sugiura M (2007) Translation of *psbC* mRNAs starts from the downstream GUG, not the upstream AUG, and requires the extended Shine-Dalgarno sequence in tobacco chloroplasts. *Plant Cell Physiol* 48:1374–1378
- Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res* 29:4633–4642
- Liu T, Dai S, Wu W, Zhang R, Fan Q, Shi S, Zhou R (2013) Development and characterization of microsatellite markers for *Melastoma dodecandrum* (Melastomataceae). *Appl Plant Sci*. <https://doi.org/10.3732/apps.1200294>
- Liu Y, Nielsen M, Staerk D, Jager AK (2014) High-resolution bacterial growth inhibition profiling combined with HPLC-HRMS-SPE-NMR for identification of antibacterial constituents in Chinese plants used to treat snakebites. *J Ethnopharmacol* 155:1276–1283. <https://doi.org/10.1016/j.jep.2014.07.019>
- Lohse M, Drechsel O, Kahlau S, Bock R (2013) OrganelleGenomeDRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Res* 41:575–581. <https://doi.org/10.1093/nar/gkt289>
- Moore MJ, Dhingra A, Soltis PS, Shaw R, Farmerie WG, Foltz KM, Soltis DE (2006) Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biol* 6:17. <https://doi.org/10.1186/1471-2229-6-17>
- Morton BR (1998) Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages. *J Mol Evol* 46:449–459
- Mower JP (2009) The PREP suite: predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. *Nucleic Acids Res* 37:253–259. <https://doi.org/10.1093/nar/gkp337>
- Ng WL, Cai Y, Wu W, Zhou R (2018) The complete chloroplast genome sequence of *Melastoma candidum* (Melastomataceae). *Mitochondrial DNA Part B* 2:242–243. <https://doi.org/10.1080/23802359.2017.1318680>
- Nikiforova SV, Cavalieri D, Velasco R, Goremykin V (2013) Phylogenetic analysis of 47 chloroplast genomes clarifies the contribution of wild species to the domesticated apple maternal line. *Mol Biol Evol* 30:1751–1760. <https://doi.org/10.1093/molbev/mst092>
- Ravi V, Khurana JP, Tyagi AK, Khurana P (2008) An update on chloroplast genomes. *Plant Syst Evol* 271:101–122. <https://doi.org/10.1007/s00606-007-0608-0>
- Reginato M, Neubig KM, Majure LC, Michelangeli FA (2016) The first complete plastid genomes of Melastomataceae are highly structurally conserved. *Peer J* 4:e2715. <https://doi.org/10.7717/peerj.2715>
- Renner SS, Meyer K (2001) Melastomeae come full circle: biogeographic reconstruction and molecular clock dating. *Evolution* 55:1315–1324
- Scharff LB, Bock R (2014) Synthetic biology in plastids. *Plant J* 78:783–798. <https://doi.org/10.1111/tpj.12356>
- Sciences EboCfotCAo (1984) *Flora of China*, vol 53. Science Press, Beijing
- Shen X et al (2017) Complete chloroplast genome sequence and phylogenetic analysis of the medicinal plant *Artemisia annua*. *Molecules* 22:1330. <https://doi.org/10.3390/molecules22081330>
- Shi C, Liu Y, Huang H, Xia E-H, Zhang H-B, Gao L-Z (2013) Contradiction between plastid gene transcription and function due to complex posttranscriptional splicing: an exemplary study of *ycf15* function and evolution in angiosperms. *PLoS ONE* 8:e59620. <https://doi.org/10.1371/journal.pone.0059620>
- Smith DR, Keeling PJ (2015) Mitochondrial and plastid genome architecture: reoccurring themes, but significant differences at the extremes. *Proc Natl Acad Sci USA* 112:10177–10184. <https://doi.org/10.1073/pnas.1422049112>
- Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, Greiner S (2017) GeSeq—versatile and accurate annotation of organelle genomes. *Nucleic Acids Res* 45:6–11. <https://doi.org/10.1093/nar/gkx391>
- Wang RJ, Cheng CL, Chang CC, Wu CL, Su TM, Chaw SM (2008) Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. *BMC Evol Biol* 8:36. <https://doi.org/10.1186/1471-2148-8-36>

- Wicke S, Schneeweiss GM, dePamphilis CW, Muller KF, Quandt D (2011) The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Mol Biol* 76:273–297. <https://doi.org/10.1007/s11103-011-9762-4>
- Wu FH et al (2010) Complete chloroplast genome of *Oncidium Gower Ramsey* and evaluation of molecular markers for identification and breeding in *Oncidiinae*. *BMC Plant Biol* 10:68. <https://doi.org/10.1186/1471-2229-10-68>
- Yang M et al (2010) The complete chloroplast genome sequence of date palm (*Phoenix dactylifera* L). *PLoS ONE* 5:e12762. <https://doi.org/10.1371/journal.pone.0012762>
- Yang GX, Zhang RZ, Lou B, Cheng KJ, Xiong J, Hu JF (2014) Chemical constituents from *Melastoma dodecandrum* and their inhibitory activity on interleukin-8 production in HT-29 cells. *Nat Prod Res* 28:1383–1387. <https://doi.org/10.1080/14786419.2014.903480>
- Zhao Y et al (2015) The complete chloroplast genome provides insight into the evolution and polymorphism of *Panax ginseng*. *Front Plant Sci* 5:696
- Zou P et al (2017) Similar morphologies but different origins: hybrid status of two more semi-creeping taxa of *Melastoma*. *Front Plant Sci* 8:673. <https://doi.org/10.3389/fpls.2017.00673>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.