The first complete plastid genomes of Melastomataceae are highly structurally conserved

Marcelo Reginato Corresp., [1] , **Kurt M Neubig** [2] , **Lucas C Majure** [3] , **Fabian A Michelangeli** [1]

[1] Institute of Systematic Botany, The New York Botanical Garden, Bronx, New York, United States

[2] Department of Plant Biology, University of Carbondale, Carbondale, Illinois, United States

[3] Department of Research, Conservation and Collections, Desert Botanical Garden, Phoenix, Arizona, United States

Corresponding Author: Marcelo Reginato
Email address: reginatobio@yahoo.com.br

Background. In the past three decades, several studies have predominantly relied on a small sample of the plastome to infer deep phylogenetic relationships in the species-rich Melastomataceae. Here, we report the first full plastid sequences of this family, compare general features of the sampled plastomes to other sequenced Myrtales, and survey the plastomes for highly informative regions for phylogenetics.

Methods. Genome skimming was performed for 16 species spread across the Melastomataceae. Plastomes were assembled, annotated and compared to eight sequenced plastids in the Myrtales. Phylogenetic inference was performed using Maximum Likelihood on six different data sets, where putative biases were taken into account. Summary statistics were generated for all introns and intergenic spacers with suitable size for PCR amplification and used to rank the markers by phylogenetic information.

Results. The majority of the plastomes sampled are conserved in gene content and order, as well as in sequence length and GC content within plastid regions and sequence classes. Departures include the putative presence of *rps16* and *rpl2* pseudogenes in some plastomes. Phylogenetic analyses of the majority of the schemes analyzed resulted in the same topology with high values of bootstrap support. Although there is still uncertainty in some relationships, in the highest supported topologies only two nodes received bootstrap values lower than 95%.

Discussion. Melastomataceae plastomes are no exception for the general patterns observed in the genomic structure of land plant chloroplasts, being highly conserved and structurally similar to most other Myrtales. Despite the fact that the full plastome phylogeny shares most of the clades with the previously widely used and reduced data set, some changes are still observed and bootstrap support is higher. The plastome data set presented here is a step towards phylogenomic analyses in the Melastomataceae and will be a useful resource for future studies.

The first complete plastid genomes of Melastomataceae are highly structurally conserved

Marcelo Reginato[1], Kurt M. Neubig[2], Lucas C. Majure[3], Fabian A. Michelangeli[1]

[1] Institute of Systematic Botany, The New York Botanical Garden, Bronx, NY, USA

[2] Dept. of Plant Biology, Southern Illinois University of Carbondale, Carbondale, Illinois 62901

[3] Dept. of Research, Conservation and Collections, Desert Botanical Garden, Phoenix, AZ, USA

Corresponding author:

Marcelo Reginato[1]

Email address: reginatobio@yahoo.com.br

Abstract

Background. In the past three decades, several studies have predominantly relied on a small sample of the plastome to infer deep phylogenetic relationships in the species-rich Melastomataceae. Here, we report the first full plastid sequences of this family, compare general features of the sampled plastomes to other sequenced Myrtales, and survey the plastomes for highly informative regions for phylogenetics.

Methods. Genome skimming was performed for 16 species spread across the Melastomataceae. Plastomes were assembled, annotated and compared to eight sequenced plastids in the Myrtales. Phylogenetic inference was performed using Maximum Likelihood on six different data sets, where putative biases were taken into account. Summary statistics were generated for all introns and intergenic spacers with suitable size for PCR amplification and used to rank the markers by phylogenetic information.

Results. The majority of the plastomes sampled are conserved in gene content and order, as well as in sequence length and GC content within plastid regions and sequence classes. Departures include the putative presence of *rps16* and *rpl2* pseudogenes in some plastomes. Phylogenetic analyses of the majority of the schemes analyzed resulted in the same topology with high values of bootstrap support. Although there is still uncertainty in some relationships, in the highest supported topologies only two nodes received bootstrap values lower than 95%.

Discussion. Melastomataceae plastomes are no exception for the general patterns observed in the genomic structure of land plant chloroplasts, being highly conserved and structurally similar to most other Myrtales. Despite the fact that the full plastome phylogeny shares most of the clades with the previously widely used and reduced data set, some changes are still observed and

bootstrap support is higher. The plastome data set presented here is a step towards phylogenomic analyses in the Melastomataceae and will be a useful resource for future studies.

Keywords Chloroplast, Melastomataceae, Myrtales, NGS, phylogenomics, plastome, genome skimming

Introduction

The Melastomataceae Juss. has over 5000 species distributed predominantly across the tropical regions. The observed levels of diversity, endemism or abundance of its members across different habitats make the family an important ecological group, as well as an excellent model for a variety evolutionary studies. The Melastomataceae belong in the Myrtales, where it is sister to the small CAP clade (Crypteroniaceae, Alzateaceae and Penaeaceae), which all together form a clade sister to Myrtaceae + Vochysiaceae (Berger et al., 2015). Plastid markers along with the nuclear ribosomal spacers (nrETS and nrITS) have been the major, and very often the exclusive, source of phylogenetic information in the family. Melastomataceae debut in molecular phylogenies was in a Myrtales-focused study, based on a partial amino acid sequence of the *rbcS* gene (Martin & Dowd, 1986). This study was followed by a more comprehensive nucleotide-based phylogeny, where the plastid *rbcL* gene was analyzed (Conti, Litt & Systma, 1996). The first Melastomataceae-wide phylogeny used a plastid data set including the *rbcL* and *ndhF* genes plus the *rpl16* intron (Clausing & Renner, 2001). This plastid data set is still the most employed source of information in studies focusing on generic relationships across the family (Fritsch et al., 2004; Renner, 2004; Amorim, Goldenberg & Michelangeli, 2009; Michelangeli et al., 2011; Goldenberg et al., 2012; Michelangeli, Ulloa & Sosa, 2014; Goldenberg et al., 2015; Zeng et al.,

2016). Phylogenetic studies within lower lineages of Melastomataceae have predominantly used the plastid spacers *accD-psaI*, *atpF-atpH*, *psbK-psbI*, and *trnS-trnG*, along with the ribosomal spacers nrETS and nrITS (Bécquer-Granados et al., 2008; Reginato, Michelangeli & Goldenberg, 2010; Kriebel, Michelangeli & Kelly, 2015, Reginato & Michelangeli, 2016). Recently, the latter data set has also been used at deeper level studies (Michelangeli et al., 2013; Rocha et al., 2016).

Family-wide phylogenetic studies based on plastid markers have uncovered major relationships in the Melastomataceae, with several implications to the classification and evolutionary understanding in the family. Early studies have consolidated the sister relationship of Olisbeoideae and the remaining Melastomataceae, settling on the currently accepted family circumscription (Conti, Litt & Systma, 1996; APG, 1998; but see Clausing & Renner, 2001 for a different perspective). Latter studies focused in some tribal re-arrangements (Fritsch et al., 2004; Penneys et al., 2010, Michelangeli et al., 2011), generic placement (Amorim, Goldenberg & Michelangeli, 2009; Goldenberg et al., 2012; Michelangeli, Ulloa & Sosa, 2014; Goldenberg et al., 2015; Kriebel, 2016; Rocha et al., 2016; Zeng et al., 2016), phylogenetic evaluation of higher species-rich lineages (Michelangeli et al., 2004, Stone, 2006; Goldenberg et al., 2008; Martin et al., 2008; Michelangeli et al., 2008; Michelangeli et al., 2013), and lower taxon phylogenies (Bécquer-Granados et al., 2008; Reginato, Michelangeli & Goldenberg, 2010; Penneys, 2013; Kriebel, Michelangeli & Kelly, 2015; Gamba-Moreno & Almeda, 2014; Majure et al., 2015; Reginato & Michelangeli, 2016). Even in family-wide phylogenies, the level of variation across these few sampled plastid markers is unsatisfactory, as evidenced by low statistical support among many relationships in different published analyses. This issue becomes more prominent in phylogenetic analyses of lineages within Melastomataceae, where the plastid phylogeny is

overall weakly supported, and concatenated results tend to be dominated by the more variable nuclear ribosomal data (Reginato, Michelangeli & Goldenberg, 2010; Reginato & Michelangeli 2016).

Phylogenomic studies are sparse in the Myrtales and absent in the Melastomataceae. Currently, there are 54 full plastids of Myrtales on the NCBI database, covering three out of the nine families in the order (Lythraceae, Myrtaceae and Onagraceae). Full plastomes can potentially improve hypotheses of phylogenetic relationships within the family, as well as in the Myrtales, and provide basic information for other aspects of molecular biology (e.g., DNA barcoding, plastome evolution, development of molecular markers). Here, we present the first complete plastid genomes in the Melastomataceae, covering 16 species spread across the family. The objectives of this study are to describe the structure of the sampled plastomes; compare main features of the plastomes within the family and to other available Myrtales plastomes; and survey the plastomes for highly informative phylogenetic markers for future use.

Material and methods

Taxon sampling, DNA extraction and sequencing

Genome skimming was performed for 16 species of Melastomataceae. Sampling was based on previous family wide phylogenetic studies (Michelangeli et al., 2014; Goldenberg et al., 2015), where each sample belongs to a different major lineage of the family, either with a formal tribe status or not. Voucher information along with GenBank accession codes are presented in Table 1. Total genomic DNA was isolated from silica-dried tissue using the Qiagen DNAeasy plant mini-kit (Qiagen, Valencia, CA) following the protocol suggested by Alexander et al. (2007) or used a modified CTAB extraction where the aqueous supernatant was silica-column

purified (Neubig et al., 2014). Total DNA samples were quantified using a NanoDrop Spectrophotometer (Thermo Scientific, Waltham, MA, USA) or Qubit 2.0 (Invitrogen, Carlsbad, CA, USA). Total genomic libraries and barcoding was performed at Cold Spring Harbor Laboratories or at Rapid Genomics (Gainesville, FL) for sequencing on an Illumina HiSeq2000 platform (Illumina, Inc., San Diego, CA, USA).

Plastid genome assembly and annotation

Total reads number yielded was on average ca. 11.5 Gb per sample (s.d.= 6 Gb). Paired reads were imported into Geneious 7.1 (Biomatters Ltd., Auckland, New Zealand), trimmed by quality (at 0.05 probability) and *de novo* assembled (Geneious Assembler, "low sensitivity" option, default settings). Filtered assembled contigs (length > 1 kb) were blasted against the *Eucalyptus polybractea* plastome (NC022393). The identified plastid contigs were then reference assembled against the *E. polybractea* plastome in order to generate a single contig to construct the circular maps. Eventual short gaps were filled by iteratively mapping the total paired reads against the contig ends. Plastid annotation was performed in Geneious 7.1 with *Arabidopsis thaliana* (NC000932) and *Eucalyptus polybractea* (NC022393) as references. Graphical representations of the plastid circular and linear maps were generated with OGDRAW (Lohse et al., 2013) and the R package genoPlotR (R Core Team, 2016; Guy, Kultima & Andersson, 2010).

Plastome structure, gene content, and general characteristics of the plastid genome were compared among the 16 Melastomataceae plastomes and to eight published plastomes of Myrtales, covering all families in this order available on the NCBI website. The Myrtales plastomes included one species in the Lythraceae (*Lagerstroemia fauriei* - NC029808), one

Onagraceae (*Oenothera grandiflora* - NC029211) and six Myrtaceae (*Allosyncarpia ternata* - NC022413; *Angophora costata* - NC022412; *Corymbia gummifera* - NC022407; *Eucalyptus polybractea* - NC022393; *Eugenia uniflora* - NC027744; and *Stockwellia quadrifida* - NC022414).

Phylogenetic analyses

Three major data sets were generated for phylogenetic inference. The first included the non-coding regions (ncs data set), the second included 78 protein-coding genes (cds data set), and the third consisted of fully assembled plastomes (full data set). In all data sets one of the IR sequences was removed to reduce overrepresentation of duplicated sequences. Full plastids were aligned with MAFFT v. 7 using the FFT-NS-i x 1000 strategy (Katoh, 2013). Coding sequences were extracted from the full alignment, resulting in the cds and ncs data sets. Each gene in the cds data set was re-aligned using its translation under the same strategy of the full data set and then concatenated. Given that phylogenetic inference might be biased by poorly aligned regions with ambiguous homology, heterogeneous rates of substitution in the different codon positions, synonymous substitutions in Arginine, Leucine and Serine codons, among others (Misof & Misof, 2009; Cox et al., 2014), we further divided the three major data sets into six different schemes where we attempted to circumvent those issues. Poorly aligned regions of the ncs data set were removed using aliscore.pl with the -N and -r options (Misof & Misof, 2009), and in the cds data set; all codons coding for Arginine, Leucine and Serine were ambiguated. Thus, the final six schemes included: 1. all ncs data set (ncs); 2. ncs data set without poorly aligned sites (ncs filtered); 3. all cds data set (cds); 4. cds with A, L and S codons ambiguated (cds ambiguated); 5. translated cds (protein); 6. ncs filtered plus all cds non-ambiguated (full). Additionally, in order to carry out a more objective comparison with previous phylogenetic

148 hypotheses, we also analyzed a reduced data set that included only the three more commonly
149 used markers for family wide phylogenies in the Melastomataceae (*ndhF* and *rbcL* genes along
150 with the *rpl16* intron, concatenated).

151 Phylogenetic inference for all schemes was performed using Maximum Likelihood
152 implemented in RAxML 8.2.4 (Stamatakis, 2014). The GTR+G model was employed for all
153 nucleotide data and the PROT+G model for the protein sequences. Support was estimated
154 through 1000 bootstrap replicates. Protein-coding sequences were partitioned by codon position
155 in all schemes, while no partitioning was employed for the non-coding regions.

156 **Phylogenetic informative regions**

157 In order to identify and rank highly phylogenetically informative regions in the
158 Melastomataceae plastomes, all introns (19) and variable intergenic spacers with suitable size for
159 PCR amplification (22) were selected and compared. Each individual marker was aligned with
160 MAFFT (FFT-NS-i x 1000 strategy), and its Maximum likelihood tree inferred with RAxML
161 (not partitioned, GTR+G model, 100 bootstrap replicates). For each marker, we report the
162 number of variable sites, number of parsimony informative sites, mean sequence distance (under
163 K80 model), alignment length, mean sequence length, mean bootstrap support and distance to the
164 full scheme plastid tree (RF distance; Robinson & Foulds, 1981). The metrics were retrieved
165 using functions of the R packages ape and phangorn (Paradis, Claude & Strimmer, 2004;
166 Schliep, 2011). Markers were ranked by phylogenetic information using a weighted mean of
167 relative values of the following metrics: number of variable sites (weight=1), mean bootstrap
168 (weight=2) and distance to the full plastid tree (weight=3). For the top 10 markers identified in
169 the previous step, we designed primer pairs for PCR amplification. Primers flanking the target

regions were designed with Primer3, using the default settings (Rozen & Skaletsky, 2000). All metrics reported, as well primer design, considered only the ingroup (the 16 Melastomataceae plastids).

Results

Plastome structure

All plastomes have a quadripartite organization, with one large single copy region (LCS), one small single copy (SSC) and two inverted repeats (IRs). A circular map of the *Miconia dodecandra* plastome is presented in Figure 1 and linear maps of all Melastomataceae plastomes in Figure 2. Sequence depth ranged from 42 to 705 (mean = 289) and plastome length from 153,311 to 157,216 bp (mean = 155,806 pb). Sequence length and GC content of the different regions across the Melastomataceae plastomes are presented in Table 2. Overall, GC content is similar across species within the same plastid region, while the LSC regions has the greatest standard deviation in sequence length (s.d. = 616 bp), followed by IR (s.d. = 250 bp) and the SSC (s.d. = 126 bp).

Most plastomes have 84 protein-coding genes (CDS), 37 transfer RNA (tRNA) and 8 ribosomal (rRNA), totaling 129 genes (including duplicates and *ycf1*, *ycf2*, *ycf3* and *ycf4*). Among the duplicated genes in the IR, there are 6 CDS, 7 tRNA, and 4 rRNA. As for the plastid regions, GC content is similar across different species within the same sequence class (CDS, tRNA, rRNA, intron and intergenic spacers), whereas the greatest variation in sequence length is observed across intergenic spacers (s.d. = 617 bp). A comparative summary of length and GC content in the different sequence classes across the Melastomataceae plastomes is given in Table

3. In the majority of the species sampled, gene content and order is similar to other Myrtales plastids, such as *Lagerstroemia fauriei* (NC029808) and *Eucalyptus polybractea* (NC022393). The exceptions are *rps16* and *rpl2*, which are putative pseudogenes in some plastids. The former seems to have been pseudogenized in *Graffenrieda moritziana* and *Pterogastra divaricata* (where the first exon is absent) and in *Salpinga margaritacea* (with several insertions changing the reading frame in the second exon); the second copy of *rpl2* gene (in the IRB) is likely a pseudogene in *Salpinga margaritacea* due to a shift in the IRB-LSC boundary in that plastid, which resulted in the loss of the second exon. Additionally, some variation is observed in all region boundaries across the Melastomataceae plastomes. The LSC-IRA boundary is located in the *rps19* gene in most species, except in *S. margaritacea* where it is located in the intron of the *rpl2* gene; the IRA-SSC boundary is located in the overlapping ψ*ycf1* and *ndhF*; the SSC-IRB in the *ycf1*; and the IRB-LSC in the *rpl2-trnH* spacer or in the *trnH* gene. Introns are found in 17 genes in all Melastomataceae plastomes, including six tRNA genes and 11 protein-coding genes, from which three have two introns (*clpP*, *rps12* and *ycf3*). A comparison of the number of genes, regions and plastome length of one Melastomataceae (*M. dodecandra*) and eight Myrtales plastids is presented in Table 4. The sequence length of the full plastome and its regions in the Melastomataceae sampled here are in the range observed for other Myrtales.

Phylogenetic analyses

The majority of the six analytical schemes recovered the same topology (Figures 2 and 3B). The only exception was the "all non-coding" scheme (i.e., the full non-coding regions without filtering of dubiously aligned base pairs), where *Blakea* + *Opistocentra*, *Triolena* + *Merianthera* and *Rhynchanthera* assume a different position (Figure 3A). Pairwise tree distances among all schemes are depicted in Figure 3C, and all Maximum Likelihood trees with bootstrap

support values are given in the Supplementary Figure S1. Bootstrap support is highest in the "full" and "cds" schemes and lower in the "protein" and "all non-coding" schemes (Figure 3D). In the highest supported topologies, there are only two nodes with bootstrap values lower than 95, and those involve the relationship disagreements between the two alternate topologies (Figure 3A-B). While filtering the non-coding poorly aligned sites improved bootstrap support and also changed the topology ("ncs" vs. "ncs filtered", Figure 3), ambiguating common amino acids in the coding sequences did not have any apparent effect in the topology or support values ("cds" vs. "cds ambiguated"; Figure 3D).

The commonly used plastid data set in previous family-wide studies (*rbcL*, *ndhF* and *rpl16* intron) also resulted in a different topology from the "full" scheme, although with most clades in common (supplementary Figure S2). Disagreements involved the position of *Allomaieta*, *Trioleta* + *Merianthera*, *Blakea* + *Opisthocentra*, and *Rhynchanthera*; these disagreements manifest in nodes of low bootstrap support where, in the reduced data set, they range from 24 to 100 (mean = 73).

Phylogenetically informative regions

Summary statistics for all intron and intergenic spacers with suitable size for PCR amplification are presented in Supplementary Table S1. A list of the top 10 markers ranked by phylogenetic information, taking into account topological distance to the tree based on the "full" scheme (Figure 2), mean bootstrap support and number of variable sites is given in Table 5, and the full list is available in supplementary Table S1. All single marker phylogenies presented some disagreement to the tree based on the "full" scheme (RF tree distance ranging from 4 to 22). Bootstrap support ranged from 26 to 82 (mean = 63) and number of variable sites from 12 to

507 (mean = 224). Primer pair sequences for PCR amplification are provided for the top 5 markers in Table 6.

Discussion

Plastid genomes of higher plants are of relatively small size, ranging from 115 to 165 kb in most groups, with an average of 90 protein-coding genes across most land plants (Ravi et al., 2007, Wicke et al., 2011). In general, the quadripartite organization, gene content and order are conserved, and GC content is usually stable within plastid regions and sequence classes (Ravi et al., 2007, Wicke et al., 2011). Melastomataceae plastomes are no exception for these patterns, being highly conserved and structurally similar to most other Myrtales, as well as to an ordinary angiosperm plastome. Melastomataceae plastomes' mean length (156 kb) is closer to the upper bound observed across most plants (165 kb), while the number of genes and GC content are around the average (90 genes, GC = 37%; Ravi et al., 2007). High conservation in genomic structure of plastomes among the Myrtales has been previously suggested (Gu et al., 2016) and is extended here to include Melastomataceae. The greatest variation in sequence length among different region classes in Melastomataceae are observed in the intergenic spacers, which is also another general pattern in plastomes (Ravi et al., 2007, Gu et al., 2016). Additionally, the boundaries of the IRs vary, as observed in some Myrtales and other groups (Bayly et al., 2013).

Conservation in gene order, content and virtual lack of recombination make the plastome a useful tool for plant phylogenetic studies (Ravi et al. 2008). An updated comprehensive phylogenetic hypothesis for the entire Melastomataceae is overdue, and full plastid sequences would contribute greatly to such an endeavor. Additionally, as sampling increases in the Myrtales, full plastids also might help to narrow down phylogenetic uncertainty in the Myrtales

(e.g., Combretaceae position, Berger et al., 2015). Despite the fact that the full plastome phylogeny recovered here shares most of the clades with the widely used "*rbcL* + *ndhF* + *rpl16*" tree, some changes are still observed and bootstrap support is higher. A more conclusive account on the extent of such changes will require more taxa to be sampled.

Here, we provide a list of potentially highly informative plastid markers for Melastomataceae. We acknowledge that the information descriptors employed are very sensitive to the taxa under analysis. Nonetheless, this ranked list can be used as guidance for sampling design of future studies, whereas the new family specific primers will increase the plastid options for Sanger sequencing-based phylogenies. There has been some debate as to whether the availability of full plastome sequencing (and other NGS tools) would render Sanger sequencing obsolete (Hert et al., 2008). Here we show that a full plastome phylogeny is an improvement on single or few plastid loci phylogenies, especially on the level of statistical support. However, considering scalability, computational complexity and budget limitations, a comprehensive NGS-based phylogeny for the mega-diverse Melastomataceae might not be achieved in the short term. Nonetheless, an expanded full plastome data set along with the more abundant Sanger-based sequences available, could be coupled in future studies. A hybrid NGS and Sanger sequencing approach has been employed for other groups (Xi et al., 2012; Leaché et al. 2014; Gardner et al., 2016), and could help clarifying the backbone of a comprehensive Melastomataceae phylogeny. Recalcitrant phylogenetic backbones are a widespread and challenging phenomenon in angiosperms (Xi et al., 2012; Straub et al., 2014), and their resolution is critical to increase the confidence of ancestral state reconstructions, historical biogeographical scenarios and other evolutionary hypotheses. Although full plastomes, or an expanded sample of plastid markers, may help to improve the confidence of phylogenetic

relationships within the Melastomataceae, we also recognize the need of parallel sampling of additional independent genealogies (i.e., nuclear and mitochondrial genomes) for further refinement in the Melastomataceae tree.

Acknowledgments

We thank all collaborators who kindly provided tissue samples for this study. We also thank Kenneth Karol for helpful insights on assembling procedures. We thank Rapid Genomics (Gainesville, FL) for sequencing services.

Funding

This study was supported by the National Science Foundation (DEB-0818399 and DEB-1343612).

References

Alexander PJ, Rajanikanth G, Bacon CD, Bailey CD. 2007. Rapid inexpensive recovery of high quality plant DNA using a reciprocating saw and silica-based columns. *Molecular Ecology Notes* **7:** 5–9.

Amorim AM, Goldenberg R, Michelangeli FA. 2009. A new species of *Physeterostemon* (Melastomataceae) from Bahia, Brazil, with notes on the phylogeny of the genus. *Systematic Botany* **34:** 324–329.

APG (Angiosperm Phylogeny Group). 1998. An ordinal classification for the families of flowering plants. *Annals of the Missouri Botanical Garden* **85:** 531–553.

Bayly MJ, Rigault P, Spokevicius A, Ladiges PY, Ades PK, Anderson C, Bossinger G, Merchant A, Udovicic F, Woodrow IE, Tibbits J. 2013. Chloroplast genome analysis of Australian eucalypts–*Eucalyptus*, *Corymbia*, *Angophora*, *Allosyncarpia* and *Stockwellia* (Myrtaceae). *Molecular Phylogenetics and Evolution* **69:** 704–716.

Bécquer-Granados ER, Neubig KM, Judd WS, Michelangeli FA, Abbott JR, Penneys DS. 2008. Preliminary molecular phylogenetic studies in *Pachyanthus* (Miconieae, Melastomataceae). *The Botanical Review* **74:** 37–52.

Berger BA, Kriebel R, Spalink D, Sytsma KJ. 2016. Divergence times, historical biogeography, and shifts in speciation rates of Myrtales. *Molecular Phylogenetics and Evolution* **95:** 116–136.

Clausing G, Renner SS. 2001. Molecular phylogenetics of Melastomataceae and Memecylaceae: implications for character evolution. *American Journal of Botany* **88:** 486–498.

Conti E, Litt A, Sytsma KJ. 1996. Circumscription of Myrtales and their relationships to other rosids: evidence from *rbcL* sequence data. *American Journal of Botany* **83:** 221–233.

Cox CJ, Li B, Foster PG, Embley TM, Civáň P. 2014. Conflicting phylogenies for early land plants are caused by composition biases among synonymous substitutions. *Systematic Biology* **63:** 272–279.

Fritsch PW, Almeda F, Renner SS, Martins AB, Cruz BC. 2004. Phylogeny and circumscription of the near-endemic Brazilian tribe Microlicieae (Melastomataceae). *American Journal of Botany* **91:** 1105–1114.

Gamba-Moreno DL, Almeda F. 2014. Systematics of the Octopleura clade of *Miconia* (Melastomataceae: Miconieae) in tropical America. *Phytotaxa* **179:** 1–174.

Gardner AG, Sessa EB, Michener P, Johnson E, Shepherd KA, Howarth DG, Jabaily RS. 2016. Utilizing next-generation sequencing to resolve the backbone of the Core Goodeniaceae and inform future taxonomic and floral form studies. *Molecular Phylogenetics and Evolution* **94:** 605–617.

Goldenberg R, Penneys DS, Almeda F, Judd WS, Michelangeli FA. 2008. Phylogeny of *Miconia* (Melastomataceae): patterns of stamen diversification in a megadiverse neotropical genus. *International Journal of Plant Sciences* **169:** 963–979.

Goldenberg R, Fraga CN, Fontana AP, Nicolas AN, Michelangeli FA. 2012. Taxonomy and phylogeny of *Merianthera* (Melastomataceae). *Taxon* **61:** 1040–1056.

Goldenberg R, Almeda F, Sosa K, Ribeiro RC, Michelangeli FA. 2015. *Rupestrea*: A new Brazilian genus of Melastomataceae, with anomalous seeds and dry indehiscent fruits. *Systematic Botany* **40:** 561–571.

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.

Guy L, Kultima JR, Andersson SG. 2010. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26:** 2334–2335.

Hert DG, Fredlake CP, Barron AE. 2008. Advantages and limitations of next-generation sequencing technologies: A comparison of electrophoresis and non-electrophoresis methods. *Electrophoresis* **29:** 4618–4626.

Katoh S. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30:** 772–780.

Kriebel R, Michelangeli FA, Kelly LM. 2015. Discovery of unusual anatomical and continuous characters in the evolutionary history of *Conostegia* (Miconieae: Melastomataceae). *Molecular Phylogenetics and Evolution* **82:** 289–313.

Kriebel R. 2016. Phylogenetic placement of the monotypic genus *Schwackaea* (Melastomeae: Melastomataceae) and the evolution of its unique fruit. *International Journal of Plant Sciences* **177:** 440–448.

Leaché, AD, Wagner P, Linkem CW, Böhme W, Papenfuss TJ, Chong RA, Lavin BR, Bauer AM, Nielsen SV, Greenbaum E, Rödel M, Schmitz A, LeBreton M, Ineich I, Chirio L, Ofori-Boateng C, Eniang EA, El Din SB, Lemmon AR, Burbrink FT. 2014. A hybrid phylogenetic–phylogenomic approach for species tree estimation in African *Agama* lizards with applications to biogeography, character evolution, and diversification. *Molecular Phylogenetics and Evolution* **79:** 215–230.

Lohse M, Drechsel O, Kahlau S, Bock R. 2013. Organellar Genome DRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Research* **41:** W575–W581.

Majure LC, Neubig KM, Skean JD, Bécquer-Granados ER, Judd WS. 2015. Evolution of the Sandpaper clade (Miconieae, Melastomataceae). *International Journal of Plant Sciences* **176:** 607–626.

Martin CV, Little DP, Goldenberg R, Michelangeli FA. 2008. A phylogenetic evaluation of *Leandra* (Miconieae, Melastomataceae): a polyphyletic genus where the seeds tell the story, not the petals. *Cladistics* **24:** 315–327.

Martin PG, Dowd JM. 1986. Phylogenetic studies using protein sequences within the order Myrtales. *Annals of the Missouri Botanical Garden* **73:** 442–448.

Michelangeli FA, Penneys DS, Giza J, Soltis D, Hils MH, Skean JD. 2004. A preliminary phylogeny of the tribe Miconieae (Melastomataceae) based on nrITS sequence data and its implications on inflorescence position. *Taxon* **53:** 279–279.

Michelangeli FA, Judd WS, Penneys DS, Skean JD, Bécquer-Granados ER, Goldenberg R, Martin CV. 2008. Multiple events of dispersal and radiation of the tribe Miconieae (Melastomataceae) in the Caribbean. *The Botanical Review* **74:** 53–77.

Michelangeli FA, Nicolas A, Morales ME, David H. 2011. Phylogenetic relationships of *Allomaieta*, *Alloneuron*, *Cyphostyla*, and *Wurdastom* (Melastomataceae) and the resurrection of the tribe Cyphostyleae. *International Journal of Plant Sciences* **172:** 1165–1178.

Michelangeli FA, Guimarães PJ, Penneys DS, Almeda F, Kriebel R. 2013. Phylogenetic relationships and distribution of new world Melastomeae (Melastomataceae). *Botanical Journal of the Linnean Society* **171:** 38–60.

Michelangeli FA, Ulloa CU, Sosa K. 2014. *Quipuanthus*, a new genus of Melastomataceae from the foothills of the Andes in Ecuador and Peru. *Systematic Botany* **39:** 533–540.

Misof B, Misof K. 2009. A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. *Systematic Biology* **58:** 21–34.

Neubig KM, Whitten WM, Abbott JR, Elliott S, Soltis DE, Soltis PS. 2014. Variables affecting DNA preservation in archival DNA specimens. In: Applequist WL, Campbell LM, eds. *DNA banking in the 21st Century: Proceedings of the U.S. workshop on DNA banking*. St. Louis: The William L. Brown Center at the Missouri Botanical Garden.

Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20:** 289–290.

Penneys DS, Michelangeli FA, Judd WS, Almeda F. 2010. Henrietteeae (Melastomataceae): a new neotropical berry-fruited tribe. *Systematic Botany* **35:** 783–800.

Penneys DS. 2013. Preliminary phylogeny of the Astronieae (Melastomataceae) based on nuclear and plastid DNA sequence data, with comments on the Philippine endemic genus, *Astrocalyx*. *Philippine Journal of Science* **142:** 159–168.

R Core Team. 2016. *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria. http://www.R-project.org/.

Ravi V, Khurana JP, Tyagi AK, Khurana P. 2008. An update on chloroplast genomes. *Plant Systematics and Evolution* **271:** 101–122.

Reginato M, Michelangeli FA, Goldenberg R. 2010. Phylogeny of *Pleiochiton* (Melastomataceae, Miconieae): total evidence. *Botanical Journal of the Linnean Society* **162:** 423–434.

Reginato M, Michelangeli FA. 2016. Untangling the phylogeny of *Leandra* s. str. (Melastomataceae, Miconieae). *Molecular Phylogenetics and Evolution* **96:** 17–32.

Renner SS. 2004. Bayesian analysis of combined chloroplast loci, using multiple calibrations, supports the recent arrival of Melastomataceae in Africa and Madagascar. *American Journal of Botany* **91:** 1427–1435.

Robinson D, Foulds LR. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences* **53:** 131–47.

Rocha MJR, Guimarães PJF, Michelangeli FA, Romero R. 2016. Phylogenetic placement and a new circumscription of *Poteranthera* (Microlicieae; Melastomataceae). *Phytotaxa* **263:** 219–232.

Rozen S, Skaletsky H. 2000. *Primer3 on the WWW for general users and for biologist programmers*. In: Misener S, Krawetz SA, eds. Methods in Molecular Biology 132: Bioinformatics methods and protocols, 365–386. New Jersey: Humana Press, Totowa, USA.

Schliep KP. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics* **27:** 592–593.

Stamatakis A. 2014. RAxML Version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30:** 1312–1313.

Stone RD. 2006. Phylogeny of major lineages in Melastomataceae, subfamily Olisbeoideae: utility of nuclear glyceraldehyde 3-phosphate dehydrogenase (*GapC*) gene sequences. *Systematic Botany* **31:** 107–121.

Straub SC, Moore MJ, Soltis PS, Soltis DE, Liston A, Livshultz T. 2014. Phylogenetic signal detection from an ancient rapid radiation: Effects of noise reduction, long-branch attraction, and model selection in crown clade Apocynaceae. *Molecular Phylogenetics and Evolution* **80:** 169–185.

Wicke S, Schneeweiss GM, Müller KF, Quandt D. 2011. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Molecular Biology* **76:** 273–297.

Xi Z, Ruhfel BR, Schaefer H, Amorim AM, Sugumaran M, Wurdack KJ, Endress PK, Matthews ML, Stevens PF, Mathews S, Davis CC. 2012. Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. *Proceedings of the National Academy of Sciences* **109:** 17519–17524.

Zeng SJ, Zou LH, Wang P, Hong WJ, Zhang GQ, Chen LJ, Zhuang XY. 2016. Preliminary phylogeny of *Fordiophyton* (Melastomataceae), with the description of two new species. *Phytotaxa* **247:** 45–61.

Figure 1

Figure 1

Map of the *Miconia dodecandra* plastid genome. Genes shown outside the outer circle are transcribed clockwise and genes inside the outer circle are transcribed counterclockwise. Genes in different functional groups are color coded following the legend. The shaded area inside the inner circle indicates the GC content, with dark shading indicating percent CG.

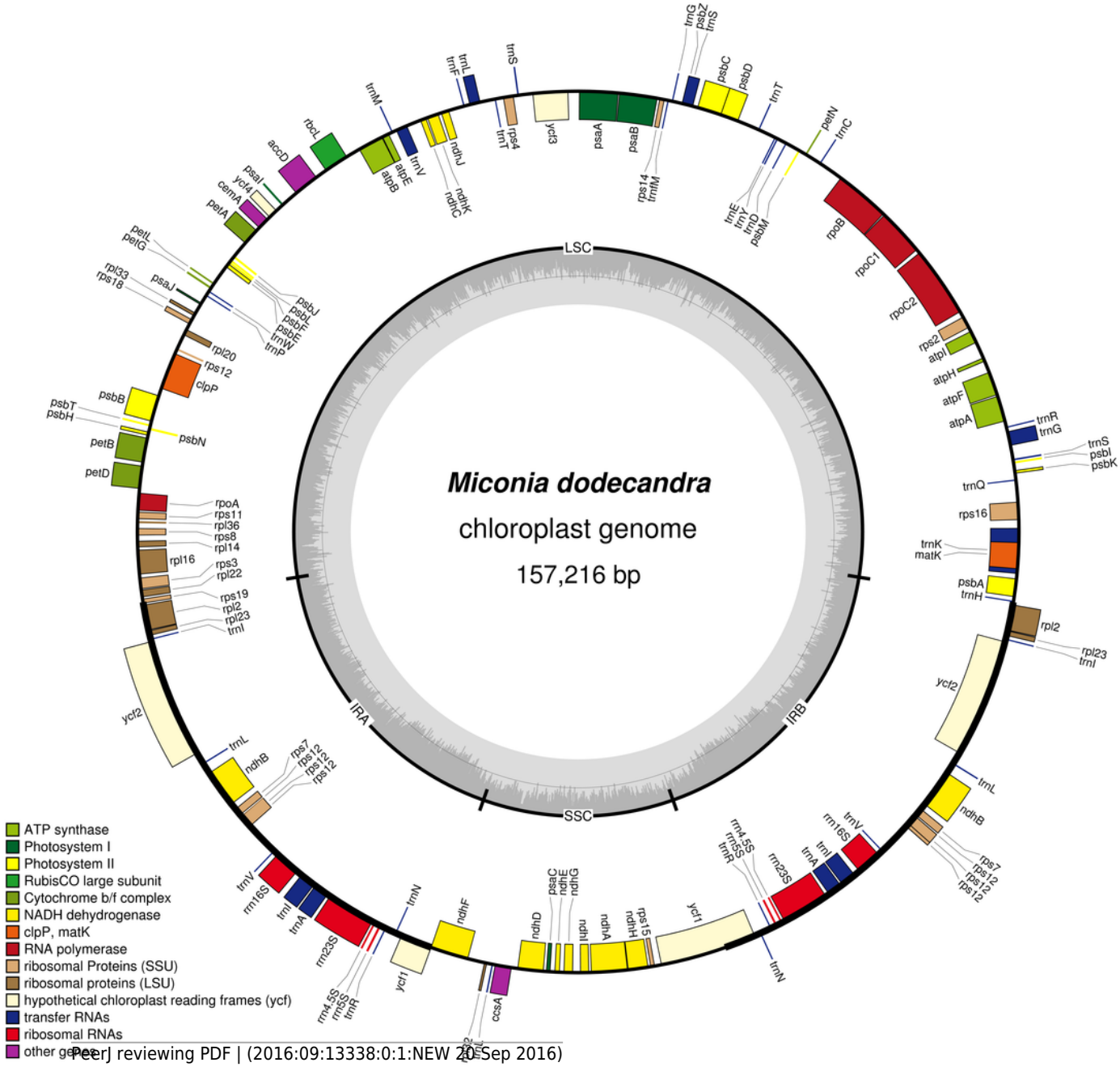

Figure 2

Figure 2

Maximum likelihood tree recovered with the full data set (left). On the right, linear plastid maps of the 16 Melastomataceae species. All genes are depicted as arrows (indicating transcription direction) and color coded following the legend of Figure 1. Gray lines link the same genes on contiguous maps. LSC (long single copy region); SSC (small single copy region); IRA (inverted repeat A); IRB (inverted repeat B).

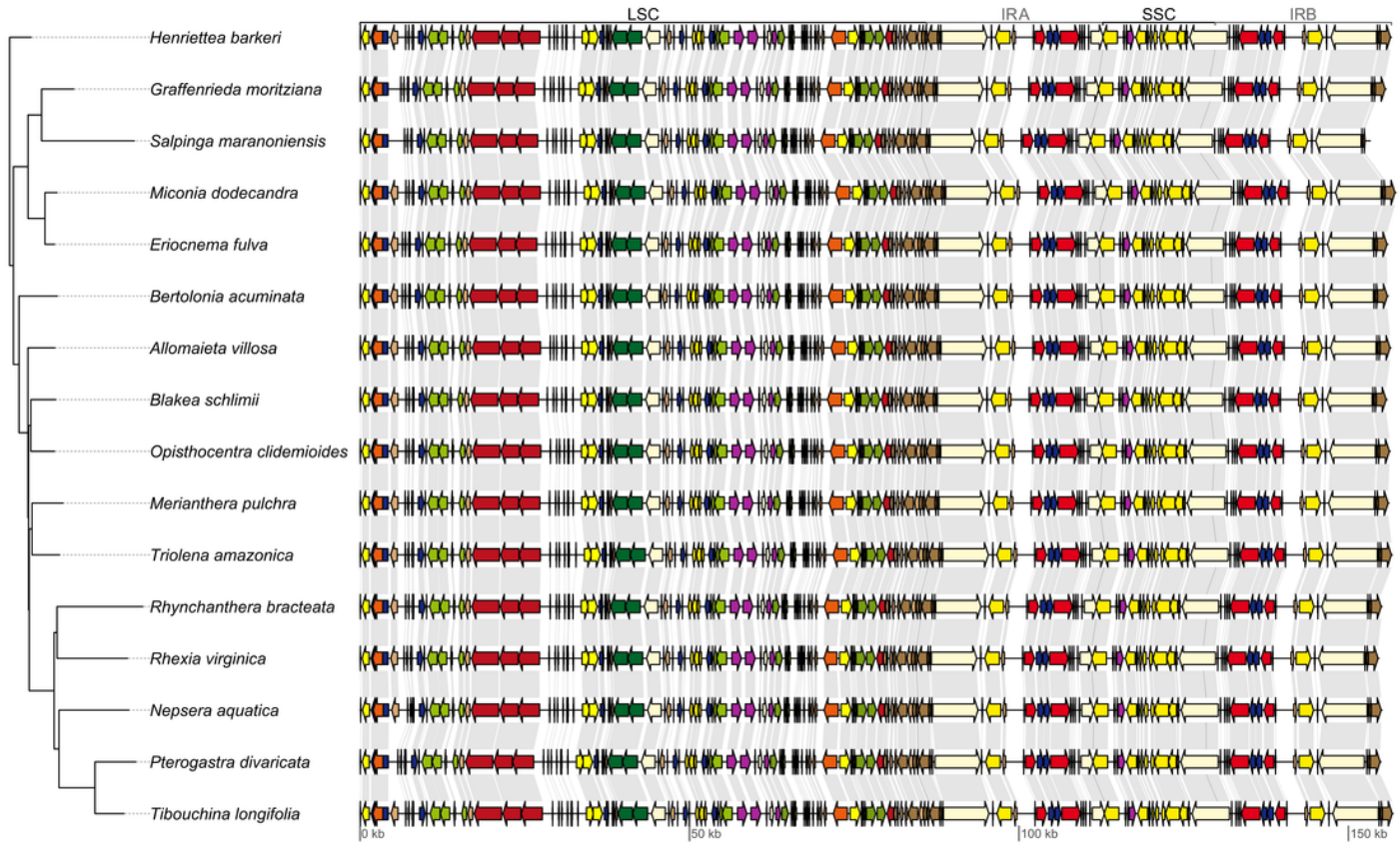

Figure 3

Figure 3

Maximum likelihood trees of the all non-coding – ncs (A) and all coding genes – cds (B) data sets. Bootstrap support is given adjacent to the nodes. C. Tree distance (RF) pairwise matrix between all six schemes analyzed. D. Mean bootstrap support of all six schemes analyzed.

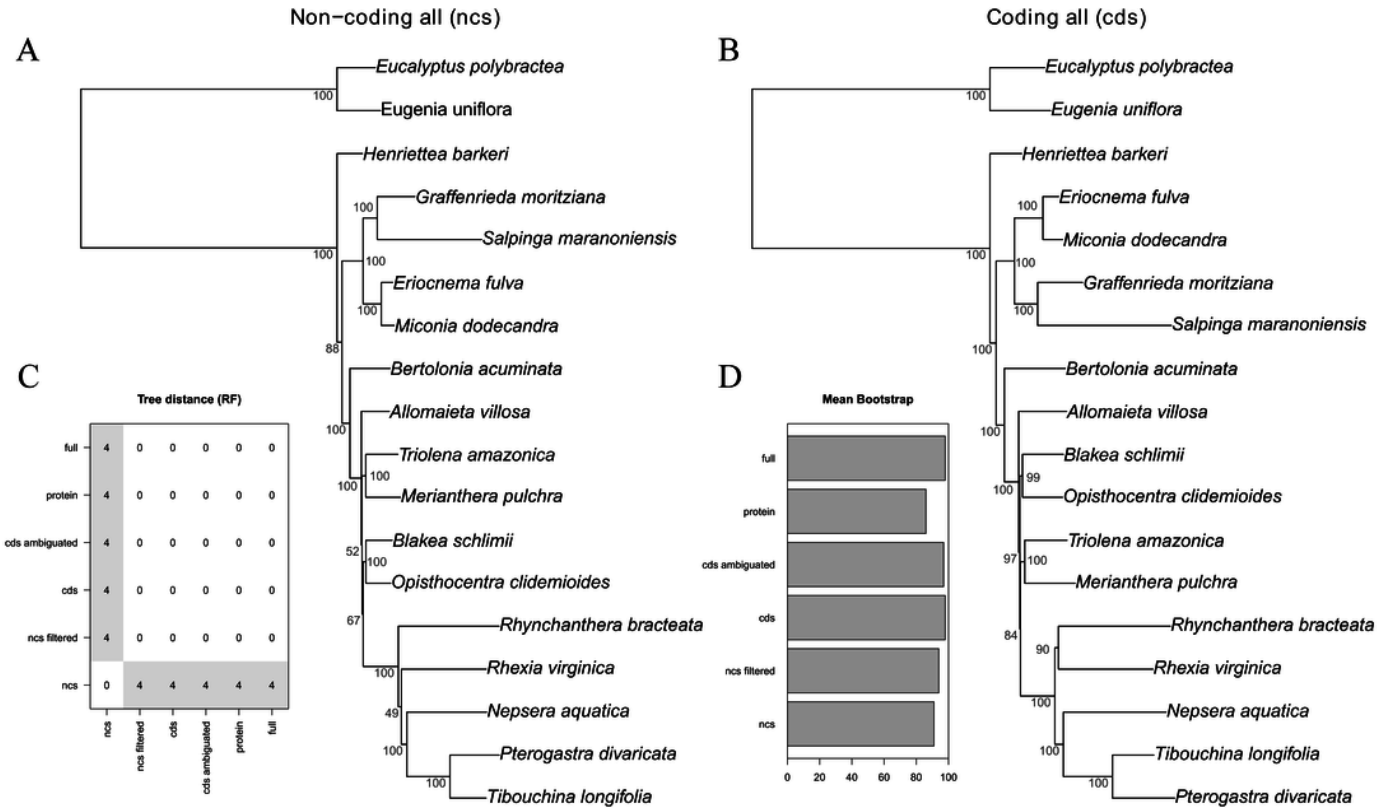

Table 1(on next page)

Table 1

Vou cher information and GenBank accessions of the chloroplast sequenced in the Melastomataceae. Informal clades are quoted.

Species	Tribe / “clade”	Genbank	Voucher	Herbarium
Allomaieta villosa (Gleason) Lozano	Cyphostyleae	KX826819	David, H. 2188	HUA, NY
Bertolonia acuminata Gardner	Bertolonieae	KX826820	Goldenberg, R. 810	NY, UPCB
Blakea schlimii (Naudin) Triana	Blakeeae	KX826821	Michelangeli, F.A. 1227	NY
Eriocnema fulva Naudin	"Eriocnema"	KX826822	Almeda, F. 8416	CAS
Graffenrieda moritziana Triana	Merianieae	KX826823	Michelangeli, F.A. 832	NY
Henriettea barkeri (Urb. & Ekman) Alain	Henrietteeae	KX826824	Ionta, G. 2029	FLAS
Merianthera pulchra Kuhlm.	"Cambessedesia"	KX826825	Goldenberg, R. 1153	NY, UPCB
Miconia dodecandra Cogn.	Miconieae	KX826826	Michelangeli, F.A. 758	NY
Nepsera aquatica (Aubl.) Naudin	"Marcetia"	KX826827	Michelangeli, F.A. 1998	NY
Opisthocentra clidemioides Hook. f.	Unplaced	KX826828	Caddah, M.K. 578	NY, UPCB
Pterogastra divaricata (Bonpl.) Naudin	Melastomeae	KX826829	Michelangeli, F.A. 540	NY
Rhexia virginica L.	Rhexieae	KX826830	Michelangeli, F.A. 1448	NY
Rhynchanthera bracteata Triana	Microlicieae	KX826831	Zenteno, F. 8801	NY
Salpinga maranoniensis Wurdack	Merianieae	KX826832	Clark, J.L. 13577	UNA
Tibouchina longifolia (Vahl) Baill.	Melastomeae	KX826833	Majure, L. 4277	FLAS
Triolena amazonica (Pilg.) Wurdack	"Triolena"	KX826834	Michelangeli, F.A. 1366	NY

1

Table 2(on next page)

Table 2

Comparison of plastid genome size and GC content across different regions in the 16 Melastomataceae species. Length (bp, total %); GC (GC content %). LSC (long single copy region); SSC (small single copy region); IR (inverted repeat); Full (full plastome).

Species	Coverage	LSC		SSC		IR		Full	
	(mean)	bp	GC	bp	GC	bp	GC	bp	GC
Allomaieta villosa	278	85915	0.347	16975	0.306	26781	0.425	156452	0.369
Bertolonia acuminata	189	85571	0.347	17008	0.308	26733	0.425	156045	0.370
Blakea schlimii	170	85370	0.349	16998	0.308	26747	0.425	155862	0.370
Eriocnema fulva	42	85431	0.348	16953	0.308	26805	0.425	155994	0.370
Graffenrieda moritziana	683	85341	0.347	16924	0.309	26734	0.425	155733	0.370
Henriettea barkeri	130	85991	0.347	17036	0.306	26750	0.425	156527	0.369
Merianthera pulchra	56	85621	0.348	17001	0.307	26773	0.424	156168	0.370
Miconia dodecandra	318	86609	0.348	16999	0.310	26804	0.425	157216	0.370
Nepsera aquatica	705	84644	0.348	17066	0.310	26700	0.426	155110	0.371
Opisthocentra clidemioides	100	85866	0.348	16942	0.309	26772	0.425	156352	0.370
Pterogastra divaricata	184	84718	0.351	17156	0.312	26537	0.425	154948	0.372
Rhexia virginica	683	84459	0.351	16924	0.311	26626	0.425	154635	0.372
Rhynchanthera bracteata	304	85093	0.347	16729	0.307	26643	0.426	155108	0.370
Salpinga maranoniensis	537	85128	0.353	16653	0.317	25765	0.428	153311	0.374
Tibouchina longifolia	195	86297	0.349	17124	0.311	26684	0.425	156789	0.371
Triolena amazonica	48	86200	0.347	16970	0.307	26741	0.425	156652	0.369

1

Table 3(on next page)

Table 3

Comparison of length and GC content across different sequence classes in the plastome of the 16 Melastomataceae species. Length (bp, total %); GC (GC content %).

Species	Protein-coding		tRNA		rRNA		Intron		Intergenic	
	bp	GC	bp	GC	bp	GC	bp	GC	bp	GC
Allomaieta villosa	80826	0.374	3348	0.497	9050	0.425	20553	0.347	42675	0.316
Bertolonia acuminata	80670	0.375	3356	0.497	9050	0.425	20437	0.347	42532	0.316
Blakea schlimii	80742	0.375	3348	0.498	9050	0.425	20541	0.347	42181	0.319
Eriocnema fulva	80628	0.375	3354	0.497	9050	0.425	20540	0.347	42422	0.318
Graffenrieda moritziana	80286	0.375	3349	0.497	9050	0.425	19691	0.347	43357	0.317
Henriettea barkeri	80781	0.374	3363	0.495	9050	0.425	20571	0.347	42762	0.315
Merianthera pulchra	80751	0.375	3364	0.498	9050	0.425	20478	0.347	42525	0.318
Miconia dodecandra	80586	0.376	3354	0.498	9050	0.425	20548	0.347	43678	0.317
Nepsera aquatica	80646	0.375	3370	0.496	9050	0.425	20619	0.347	41425	0.318
Opisthocentra clidemioides	80643	0.376	3360	0.496	9050	0.425	20641	0.347	42658	0.317
Pterogastra divaricata	80427	0.377	3339	0.498	9050	0.425	19911	0.347	42221	0.318
Rhexia virginica	80466	0.377	3353	0.496	9050	0.425	20260	0.347	41506	0.319
Rhynchanthera bracteata	80415	0.375	3241	0.502	9048	0.425	20538	0.347	41866	0.317
Salpinga maranoniensis	79326	0.376	3349	0.500	9050	0.425	18991	0.347	42595	0.326
Tibouchina longifolia	80682	0.377	3348	0.497	9050	0.425	20666	0.347	43043	0.317
Triolena amazonica	80619	0.375	3337	0.496	9050	0.425	20476	0.347	43170	0.316

1

Table 4(on next page)

Table 4

Comparison of plastid genome size of one Melastomataceae species (*Miconia dodecandra*) with eight other Myrtales. Protein-coding, tRNA and rRNA (number of genes); LSC (long single copy region, length in bp), SSC (small single copy region, length in bp), IR (inverted repeat, length in bp) and Full (length in bp).

Family	Species	Coding	tRNA	rRNA	LSC	SSC	IR	Full
Melastomataceae	*Miconia dodecandra*	84	37	8	86609	16999	26804	157216
Myrtaceae	*Allosyncarpia ternata*	84	37	8	88218	18571	26402	159563
Myrtaceae	*Angophora costata*	84	37	8	88769	18773	26392	160326
Myrtaceae	*Corymbia gummifera*	84	37	8	88310	17197	27603	160713
Myrtaceae	*Eucalyptus polybractea*	84	37	8	88944	18530	26397	160268
Myrtaceae	*Eugenia uniflora*	84	37	8	87459	18318	26334	158445
Lythraceae	*Lagerstroemia fauriei*	84	37	8	83923	16933	25792	152440
Onagraceae	*Oenothera grandiflora*	84	38	8	89862	19035	28824	166545
Myrtaceae	*Stockwellia quadrifida*	84	37	8	88247	18544	26385	159561

1

2

Table 5(on next page)

Table 5

Summary statistics for the top 10 introns and intergenic spacers with suitable size for PCR amplification. Markers are ranked by phylogenetic information based on a weighed mean of relative values of number of variable sites (weight=1), mean bootstrap (weight=2) and distance to the full plastid tree (weight=3). PIS = parsimony informative sites; Tree distance = RF distance.

Marker	Bases	Aligned (bp)	Variable sites	PIS	DNA distance (mean)	Tree distance	Bootstrap (mean)
1. *trnS-trnG* spacer	780 [628,884]	1125	438 (38.9 %)	128 (11.4 %)	0.104	4	82
2. *ndhF-rpl32* spacer	898 [849,965]	1266	507 (40 %)	171 (13.5 %)	0.114	6	71
3. *trnG* intron	762 [743,790]	846	236 (27.9 %)	76 (9 %)	0.059	4	75
4. *ndhC-trnV* spacer	734 [504,821]	991	330 (33.3 %)	98 (9.9 %)	0.081	4	63
5. *ndhA* intron	1016 [939,1045]	1127	250 (22.2 %)	74 (6.6 %)	0.046	4	64
6. *trnG-atpA* spacer	641 [550,750]	895	353 (39.4 %)	136 (15.2 %)	0.114	6	65
7. *atpH-atpI* spacer	898 [638,980]	1178	323 (27.4 %)	92 (7.8 %)	0.062	8	76
8. *psbE-petL* spacer	1058 [570,1165]	1396	381 (27.3 %)	132 (9.5 %)	0.068	8	70
9. *petA-psbJ* spacer	736 [420,944]	1062	285 (26.8 %)	90 (8.5 %)	0.076	8	76
10. *trnE-trnT* spacer	842 [478,1029]	1345	406 (30.2 %)	121 (9 %)	0.089	8	63

1

2

Table 6(on next page)

Table 6

Primer pair sequences for the indentified top 5 highly informative markers across the 16 plastomes of Melastomataceae.

Marker	Primer forward (5'-3')	Primer reverse (5'-3')	T_a (°C)
1. *trnS-trnG* spacer	CACTCAGCCATCTCTCCCAA	ACCCGCTACAATGCCATTATTG	55
2. *ndhF-rpl32* spacer	AGGAAAGGACCACATACGTCG	TCCTTGCTCATTGATTTTGATCCA	55
3. *trnG* intron	GGTCCCTCGGATTTGCTTCA	GAACCCGCATCGTTAGCTTG	55
4. *ndhC-trnV* spacer	AGATGAACTCCTAGGGAATGTGA	CCGAGAAGGTCTACGGTTCG	55
5. *ndhA* intron	CGCTAGTCCAGAACCGTACA	ACCCCATGATTGGTTGATTAGTGA	55

1