Plant Syst Evol (2017) 303:1251–1270
DOI 10.1007/s00606-017-1436-5

ORIGINAL ARTICLE

Plastid genome structure and phylogenomics of Nymphaeales: conserved gene order and new insights into relationships

Michael Gruenstaeudl[1] · Lars Nauheimer[2] · Thomas Borsch[1,3,4]

Received: 16 March 2017 / Accepted: 16 June 2017 / Published online: 6 July 2017

Abstract The plastid genomes of early-diverging angiosperms were among the first land plant plastomes investigated. Despite their importance to understanding angiosperm evolution, no investigation has so far compared gene content or gene synteny of these plastid genomes with a focus on the Nymphaeales. Here, we report an evaluation and comparison of gene content, gene synteny and inverted repeat length for a set of 15 plastid genomes of early-diverging angiosperms. Seven plastid genomes of the Nymphaeales were newly sequenced for this investigation. We compare gene order and inverted repeat (IR) length across all genomes, review the gene annotations of previously published genomes, generate a multi-gene alignment of 77 plastid-encoded genes and reconstruct the phylogenetic relationships of the taxa under study. Our results show that gene content and synteny are highly conserved across early-diverging angiosperms: All species analyzed display complete gene synteny when accounting for expansions and contractions of the IRs. This conservation was initially obscured by ambiguous and potentially incorrect gene annotations in previously published genomes. We also report the presence of intact open reading frames across all taxa analyzed. The multi-gene phylogeny displays maximum support for the families Cabombaceae and Hydatellaceae, but no support for a clade of all Nymphaeaceae. It further indicates that the genus *Victoria* is embedded within *Nymphaea*. Plastid genomes of *Trithuria* were found to deviate by numerous substitutions and length changes in the IRs. Phylogenetic analyses further indicate that a previously published plastome named *Nymphaea mexicana* falls into a clade of *N. odorata* and should be re-evaluated.

Keywords Early-diverging angiosperms · Gene annotations · Gene order · Inverted repeats · Nymphaeales · Plastid genome

Introduction

A characteristic feature of the circular plastid genome is its quadripartite structure, with the presence of two large inverted repeats (IRs) that separate two single-copy regions: the large single-copy (LSC) and the small single-copy (SSC) region. The IRs can vary in size even in closely related species or, in rare cases, be entirely absent (Guisinger et al. 2011; Wicke et al. 2011). The complete size of plastid genomes typically varies between 120 and 160 kbp (Wicke et al. 2011), but can range from 85 kbp in the parasitic *Cuscuta* (McNeal et al. 2007) to 242 kbp in *Pelargonium* (Weng et al. 2014). Size differences are usually the result of expansions or

Handling editor: Marcus Koch.

Michael Gruenstaeudl and Lars Nauheimer have contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00606-017-1436-5) contains supplementary material, which is available to authorized users.

✉ Michael Gruenstaeudl
m.gruenstaeudl@fu-berlin.de

[1] Institut für Biologie, Systematische Botanik und Pflanzengeographie, Freie Universität Berlin, Berlin, Germany

[2] Australian Tropical Herbarium, Cairns, Australia

[3] Botanischer Garten und Botanisches Museum Berlin, Freie Universität Berlin, Berlin, Germany

[4] Berlin Center for Genomics in Biodiversity Research (BeGenDiv), Berlin, Germany

contractions of the IRs and, thus, the duplication of coding and non-coding regions, but are rarely due to changes in gene content, which is relatively stable in land plants (Wicke et al. 2011; Jansen and Ruhlman 2012). Strong reductions in gene content are primarily known from the plastomes of plants with parasitic (McNeal et al. 2007; Wicke et al. 2013) or mycoheterotrophic lifestyles (Logacheva et al. 2011). Considerable alterations in gene content were recently also reported from the carnivorous Lentibulariaceae (Wicke et al. 2014). By contrast, the gene order of plastid genomes in specific land plant lineages appears to be more variable. For example, extensive rearrangements were reported from the plastid genomes of *Campanula* (Campanulaceae, Haberle et al. 2008), *Trifolium* (Fabaceae, Cai et al. 2008) and *Pelargonium* (Geraniaceae, Weng et al. 2014). It is currently unclear if genomic rearrangements occur stochastically throughout the angiosperms or if certain lineages display a higher propensity for such rearrangements (Weng et al. 2014), which might be connected to idiosyncrasies in their DNA replication and repair systems (Zhang et al. 2016).

Since the first complete plastid genome sequences were applied to plant phylogenetics approximately 15 years ago, plastid phylogenomics has become widely used to resolve phylogenetic relationships among photosynthetic eukaryotes, including land plants (Ruhfel et al. 2014) and green algae (Sun et al. 2016). The increased availability of complete plastid genomes considerably increased the character base for phylogenetic analyses of disputed evolutionary relationships, including the early-diverging angiosperms. An active discussion has developed on the question if *Amborella* constitutes the sister to all angiosperms (e.g., Soltis et al. 1999; Borsch et al. 2003; Leebens-Mack et al. 2005; Müller et al. 2006; Jansen et al. 2007; Moore et al. 2007, 2011; Drew et al. 2014) or if it forms a clade with the Nymphaeales (e.g., Goremykin et al. 2003a, 2004, 2013). Even though the sequencing of complete plastid genomes has dramatically increased the number of genes under study (Goremykin et al. 2003a, 2004), a dense taxon sampling and the choice of methodology were found similarly important toward accurate phylogeny inference in early-diverging angiosperms (Jansen et al. 2007; Moore et al. 2007). In an attempt to increase taxon sampling in the Nymphaeales, two species of the family Hydatellaceae, *Trithuria inconspicua* (Goremykin et al. 2013) and *Trithuria filamentosa* (Drew et al. 2014), have recently been added to the so far sequenced plastid genomes of *Nuphar advena* (Raubeson et al. 2007) and *Nymphaea alba* (Goremykin et al. 2003a). This addition improved the representation of species of the Nymphaeales among the available plastid genomes of early-diverging angiosperms, but the family Cabombaceae has yet to be included.

The plant order Nymphaeales takes a pivotal role in the discussion on the evolutionary history of early-diverging angiosperms. A dense sampling of plastid genomes from this order would allow to test hypotheses concerning gene order and size of inverted repeats in the ancestor of all extant angiosperms. However, the phylogenetic relationships within the Nymphaeales are only partially understood. More than half of the diversity of the Nymphaeales can be found in the genus *Nymphaea* (Schneider and Williamson 1993; Borsch et al. 2008, 2011). The stem of the Nymphaeales is estimated to have diverged 108.6 Ma (±25.2 Ma), whereas the major diversifications occurred in the Eocene and the middle Miocene (Löhne et al. 2008; Iles et al. 2014). In its current circumscription, the Nymphaeales include the families Nymphaeaceae, Cabombaceae and Hydatellaceae, the latter of which were identified as part of the water lily clade only recently (Saarela et al. 2007).

The only direct comparison of plastid genomes of the Nymphaeales has been conducted by Raubeson et al. (2007), who compared the sequence similarity of genes, introns and intergenic regions in the plastomes of *Nuphar advena* and *Nymphaea alba*. They reported that more than two-thirds of the homologous regions are at least 95% identical and that only 3% of the regions are less than 70% identical. They also showed that the IR boundaries of *Nuphar* differed from those of *Nymphaea* and *Amborella* in that the *ndh*F gene spans across the IR/SSC boundary. To our knowledge, no subsequent investigation has specifically compared gene order and IR size of early-diverging angiosperms since then. The recently sequenced plastid genomes of *Trithuria* add an interesting new component to the structural diversity of the plastid genomes of early-diverging angiosperms. These genomes differ substantially in sequence length, displaying 5–20 kbp longer genome sequences than other early-diverging angiosperms, which may be indicative of major structural changes or of IR expansions.

In the present investigation, we generate complete and annotated plastid genomes of seven members of the Nymphaeales. Six of these genomes represent species that have not been sequenced before, including members of the Cabombaceae. We combine the DNA sequences of the novel plastid genomes with those of eight previously published plastid genomes of early-diverging angiosperms, generating a dataset that comprises the Nymphaeales with all major subclades and other early-diverging angiosperms (*Amborella*, Austrobaileyales). Upon careful review and, where necessary, correction of the gene annotations in each plastid genome under study, we evaluate and compare gene content, gene synteny and IR boundaries. We also assess the presence of translation initiation and translation termination codons in several open reading frames and hypothetical protein-coding genes, which are not consistently annotated across early-diverging angiosperm plastomes. Furthermore, we generate a multi-gene alignment of 77 plastid-encoded genes of the 15 plastid genomes under study and reconstruct the phylogenetic relationships between these taxa. Finally, we infer the phylogenetic placement of a previously published plastid genome that was designated as "*Nymphaea mexicana*" by the original authors, but which appears to be incorrectly determined based on a DNA sequence comparison with other records of *Nymphaea*.

Materials and methods

Taxon sampling and combination with previously published genomes

The plastid genomes of five species of the Nymphaeaceae and two species of the Cabombaceae were newly sequenced for this investigation. Specifically, we generated complete plastid genomes of the genera *Barclaya*, *Nymphaea* (*N. alba*, *N. ampla*, *N. jamesoniana*) and *Victoria* of the Nymphaeaceae and of the genera *Brasenia* and *Cabomba* of the Cabombaceae. These seven plastid genomes were combined with eight previously published plastid genomes of various early-diverging angiosperm families, generating a dataset of 15 plastid genomes that represent all currently recognized orders of early-diverging angiosperms. Species name, taxonomic position and GenBank accession number of each of the 15 records as well as sampling location and herbarium voucher information for each of the newly sequenced plastid genomes are presented as electronic supplementary material (Online Resource 1).

DNA extraction and genome sequencing

All plastid genomes sequenced for this investigation were generated from young leaves taken from live plant specimens cultivated at the Botanical Garden and Botanical Museum Berlin. Prior to DNA isolation, the edges of each leaf sample were removed with sterile razor blades, and the cut leaves were rinsed with deionized water and 70% ethanol. Total genomic DNA was isolated from 1.5 g of cleaned leaf material using the NucleoSpin Plant II kit (Macherey–Nagel, Düren, Germany). For each DNA extraction, a barcoded genomic library was constructed using the Nextera DNA library preparation kit (Illumina, San Diego, CA, USA). Finished libraries were quantified using a PicoGreen dsDNA quantitation kit (Invitrogen, Carlsbad, CA, USA). The libraries were pooled and sequenced as paired-end reads (MiSeq reagent kit, v3 chemistry, 600 cycles) on an Illumina MiSeq system at the Berlin Center for Genomics in Biodiversity Research with an insert size of 250–300 bp.

Genome assembly and annotation

A minimum of 1.6 million paired-end reads were generated per sample. Raw reads were trimmed by quality via FASTX Toolkit v.0.0.14 (Gordon 2014), using a minimum quality score of 30. Upon quality filtering, between 2.61 and 4.59% of all quality-filtered reads per sample were found to be of chloroplast origin, as they mapped to either of the two previously published plastid genomes *Nymphaea alba* (NC_006050; Goremykin et al. 2004) and *Nuphar advena* (NC_008788; Raubeson et al. 2007), which were used as reference genomes. Trimmed and quality-filtered reads were assembled de novo into contigs using Velvet v.1.2.10 (Zerbino and Birney 2008), testing a range of kmer values to optimize contig length (kmer = 33–97, in increments of 4). Back-mapping of successfully assembled reads to the complete plastid genomes indicated a mean coverage depth greater than 150 for each newly generated plastid genome, with more than 98.5% of all bases covered by a depth of 50 or greater. Individual contigs were combined manually into final assemblies with the help of the software application Geneious v.7.1.9 (Kearse et al. 2012), using the reference genomes as guides for contig position and orientation. Adjacent, overlapping contigs were combined only if less than 10% of the overlapping nucleotides were dissimilar. Minimally, between two and five contigs had to be so combined to cover half of each plastid genome. Read, contig and assembly statistics across newly generated plastid genomes were calculated via QUAST v.4.5 (Gurevich et al. 2013) and are provided as electronic supplementary material (Online Resource 2).

General genome structure and ambiguous nucleotide positions were evaluated through an additional assessment. The quadripartite structure of the final assemblies and the equality of the inverted repeats were confirmed by self-blasting each assembly using the BLAST + suite v.2.4.0 (Camacho et al. 2009). Single nucleotide ambiguities were resolved by mapping the trimmed reads against the regions with ambiguities using the short read mapper bowtie v.1.1.2 (Langmead et al. 2009) and assigning those nucleotides found by majority rule across the mapped reads.

Gene annotation of the assembled plastid genomes was carried out in a two-step procedure. First, we added raw annotations as predicted by the annotation servers DOGMA (Wyman et al. 2004) and cpGAVAS (Liu et al. 2012), selecting *Nymphaea alba* (accession NC_006050) as reference genome. Second, we inspected and curated all gene annotations and annotation names manually using Geneious. During this curating, we extracted all coding regions per genome, confirmed start and stop codons for each gene, aligned the extracted regions across all study taxa, confirmed approximate gene lengths based on their amino acid translations and reconfirmed any internal stop codons. This confirmation was carried out for all 15 input taxa, thus confirming the annotations of the newly generated as well as the previously published plastomes. The complete plastid genome sequences of all newly sequenced genomes are available from GenBank (Online Resource 1).

Inference and visualization of gene synteny

To visualize gene synteny across all 15 plastid genomes under study, the genes were listed by their location in the genome in the following partition order: LSC, inverted repeat B (IR_B), SSC, inverted repeat A (IR_A). In cases where genes have multiple exons separated by one or more coding sequences (e.g., *trn*K, *rps*12), the exons instead of the full genes were used in our synteny comparisons, with "p1" and "p2" appended to their names. Circular and linear genome maps were generated via OGDRAW v.1.2 (Lohse et al. 2013; Online Resource 3).

Multi-gene alignment and manual alignment correction

To infer the phylogenetic relationships among the newly sequenced plastid genomes, a multi-gene alignment was generated. Upon bioinformatic extraction of the full gene complement from each plastid genome under study, a gene-by-gene alignment was conducted, whereby each gene was translated into a sequence of amino acids, the amino acids aligned under the scoring matrix BLOSUM62 using MAFFT v.7.304b, and the aligned amino acid sequences back-translated to nucleotide sequences. The resulting DNA sequence alignments were manually adjusted following the rules of Löhne and Borsch (2005) using PhyDE v.0.9971 (Müller et al. 2007) to assure positional homology, following the understanding that alignment at the genome level must not be less rigorous than for individual genomic regions. Small inversions were separated in specific columns to prevent incorrect homology assessment. Upon manual alignment adjustment, the gene-wise alignments were concatenated in the same gene order as found in the actual genomes. The hypothetical protein-coding gene *ycf*1 was not included in the multi-gene alignment due to large regions of uncertain homology. Ribosomal and transfer RNA genes were also not included in the alignment given the uncertainty associated with accounting for their secondary structure (Michaud et al. 2011). Moreover, we excluded several sections of the genes *acc*D, *ndh*F, *rpo*C1 and *rpo*C2 from the multi-gene alignment due to considerable uncertainty about the correct positional homology. The length of the exclusions was hereby set to a multiple of 3 bp, ensuring preservation of the reading frame.

Phylogenetic inference and data partitioning

Phylogenetic reconstructions were performed on the complete multi-gene alignment via maximum likelihood (ML) and Bayesian (BI) phylogeny inference. Analyses via ML were conducted with RAxML v.8.2.9 (Stamatakis 2014) using the best-fitting nucleotide substitution model GTR + G + I and the thorough ML optimization option. Branch support for ML analyses was calculated via 1000 bootstrap (BS) replicates using the rapid BS algorithm (Stamatakis et al. 2008) and the same nucleotide substitution models as under tree inference. Analyses via BI were conducted with MrBayes v.3.2.5 (Ronquist and Huelsenbeck 2003) under the best-fitting nucleotide substitution model as inferred by jModeltest v.2.1.7 (Darriba et al. 2012), using four parallel Markov Chain Monte Carlo (MCMC) runs for a total 20 million generations. Independent sampling of generations and convergence of Markov chains were confirmed in Tracer v.1.6 (Rambaut et al. 2014). The initial 50% of all MCMC trees were discarded as burn-in, and post-burn-in trees were summarized as a majority rule consensus tree, with branch support given as posterior probability (PP) values. The complete multi-gene alignment as well as the optimal phylogenetic trees inferred under ML and BI are available at Zenodo (https://zenodo.org/record/377039/).

To increase the accuracy of the inference of phylogenetic tree topology, branch lengths and substitution model parameters, we conducted our phylogenetic analyses under different data partitioning strategies. Specifically, we compared the results of four different partitioning strategies under both ML and BI phylogeny inference, with each strategy applied to the multi-gene DNA alignment. First, we conducted phylogenetic analyses on an unpartitioned matrix in which the entire multi-gene DNA alignment was analyzed under the nucleotide substitution model GTR + I + G; only a single partition was analyzed under this strategy. Second, we conducted phylogenetic analyses on a partitioned matrix in which each of the 77 genes of the alignment was analyzed under its best-fitting nucleotide substitution model as inferred by jModeltest; a total of 77 partitions were analyzed under this strategy. Third, we conducted phylogenetic analyses on a partitioned matrix in which each of the three codon positions across the alignment was grouped into its own partition; a total of three partitions were analyzed under this strategy. Fourth, we conducted phylogenetic analysis on a partitioned matrix that was inferred as the best-fitting partitioning strategy via PartitionFinder2 (Lanfear et al. 2016). Specifically, the software inferred a partitioning strategy with 18 different partitions as optimal given the multi-gene DNA alignment. A list of the individual partitions and their best-fitting nucleotide substitution models of the four data partitioning strategies is given as electronic supplementary material (Online Resource 4).

Evaluation of hypothetical genes *ycf*15 and *ycf*68

To investigate the presence and the DNA sequence conservation of the hypothetical protein-coding genes *ycf*15 and *ycf*68 in the plastid genomes of early-diverging angiosperms, we extracted the sequence of *ycf*15 from the plastid genome of *Nicotiana tabacum* (accession Z00044, Shinozaki et al. 1986) and the sequence of *ycf*68 from the plastid genome of *Trithuria inconspicua* and aligned them as baits to each of the 15 plastid genome sequences under study. We then extracted the best region that our baits aligned to from each genome, saved the extracted regions in gene-specific sequence sets and aligned each sequence set using MAFFT. The resulting alignments of *ycf*15 and *ycf*68 were compared for overall length, internal stop codons and shifts in reading frames. The reading frame of the bait sequences was hereby used reference for the transcribed exons in the plastid genomes. The alignments of *ycf*15 and *ycf*68 are available as electronic supplementary material (Online Resources 5 and 6).

Test of taxon designation of GenBank record NC_024542

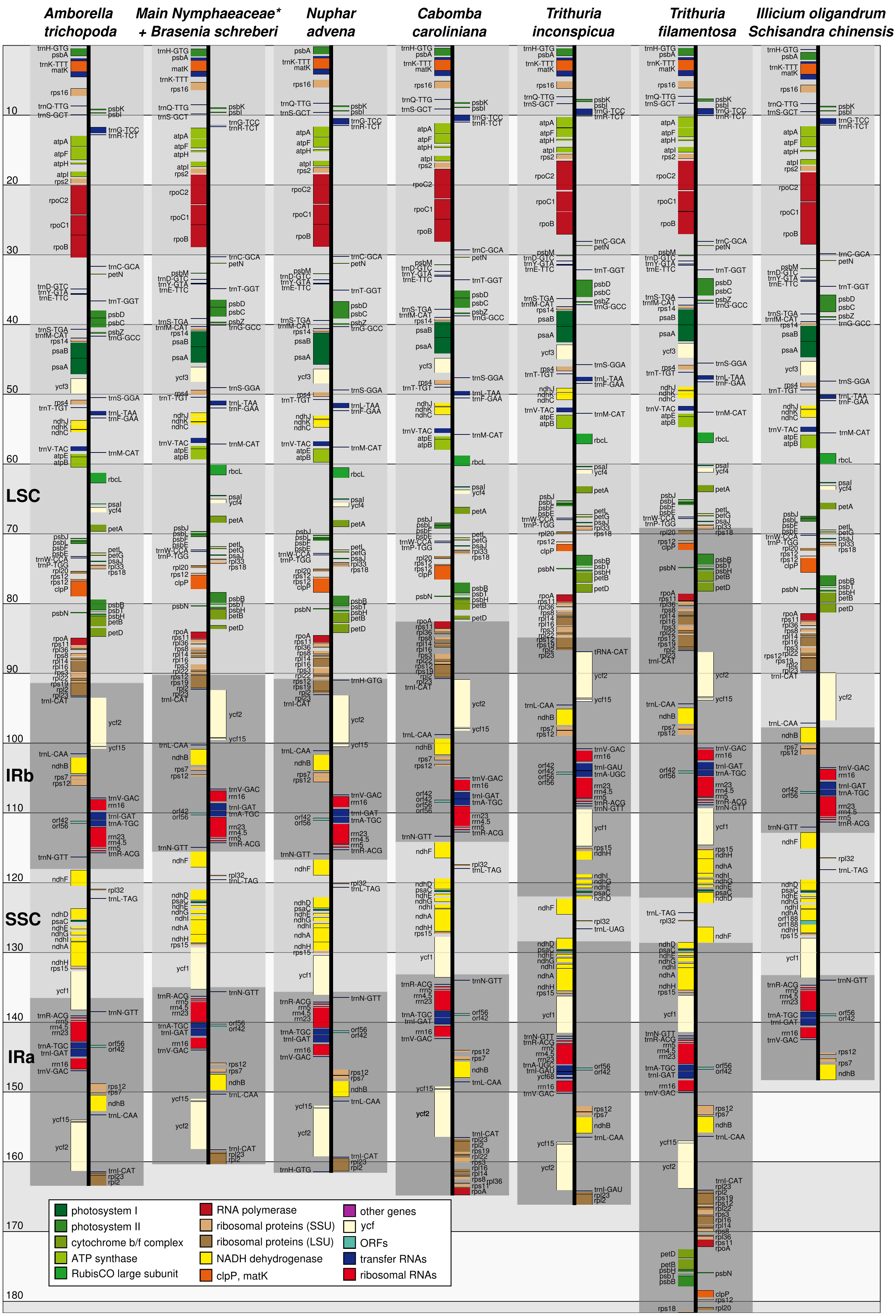
Doubts about the taxon designation of a previously published plastid genome arose given its sequence similarity to other DNA sequence records of *Nymphaea*. In a preliminary investigation, we found that the sequences of *pet*D, *rpl*16, *trn*K-*mat*K and *trn*T–*trn*F of the plastid genome with GenBank accession number NC_024542 (“*Nymphaea mexicana*”; Yang et al. 2014) were exactly identical to previously published DNA sequences of *Nymphaea odorata*. We therefore decided to evaluate the taxon designation of this plastid genome as part of a genus-wide alignment. Specifically, we extracted the intergenic plastid spacer *trn*T–*trn*F from the 15 plastid genomes under study and included them in a previously published alignment of this marker (Borsch et al. 2011). We also added DNA sequences of species of *Nymphaea* subg. *Nymphaea* (Borsch et al. 2014), of *Nuphar* (Soininen et al. 2009) and of species outside the Nymphaeales (Borsch et al. 2003) to the alignment. We then reconstructed the phylogenetic relationships among the input sequences under both ML and BI. Phylogeny inference was conducted under the same settings as the inference of relationships among the complete plastid genomes using the multi-gene alignment, except that indels were coded according to the “Simple Indel Coding” scheme (Simmons and Ochoterena 2000) using SeqState v.1.40 (Müller 2005). DNA sequences and indel coding were combined into a partitioned dataset and analyzed with unlinked parameters. The indel partition was analyzed under the binary substitution model BINGAMMAI in RAxML and under the binary character model (Lewis 2001) in MrBayes. Based on these phylogenetic inferences, we refer to the plastid genome of accession NC_024542 as “*Nymphaea* cf. *odorata*” for the remainder of this manuscript.

Results

Genome structure and length of inverted repeats

The general genome structure of the 15 plastid genomes of early-diverging angiosperms under study is highly conserved. All plastid genomes analyzed display a typical quadripartite genome organization, with IR regions separating the LSC from the SSC (Fig. 1). The complete length of the plastid genomes ranges from 147,772 bp in *Schisandra chinensis* to 180,562 bp in *Trithuria filamentosa*, with all genomes of the Nymphaeaceae displaying a length between 158,360 and 160,866 bp (Table 1). The interquartile range (IQR) of the length of the LSC is between 88,737 and 90,199 bp, of the SSC between 18,822 and 19,562 bp and of the IR between 25,144 and 26,243 bp. The length variability of the three regions hereby differs markedly, with the length of the IR region, and thus the number of genes contained in them, being most variable. Specifically, the IR regions display a greater standard deviation (SD) in sequence length (SD = 8865 bp) than the LSC (SD = 6742 bp) or the SSC (SD = 4623 bp). Genome maps of each newly sequenced plastid genome are available as electronic supplementary material (Online Resource 3).

The higher variability in length of the IR than of the LSC or the SSC across our study taxa is primarily the result of differential expansions of the IR regions in the genera *Cabomba*, *Nuphar* and *Trithuria* as well as its contractions in the two species of the Austrobaileyales under study when compared to *Amborella* (Fig. 1). In the two species of *Trithuria*, the IR_A appears to have expanded into the SSC region compared to *Amborella*, integrating the first eight adjacent genes. In *Trithuria filamentosa*, an additional expansion is found, whereby the IR_B appears to have expanded into the LSC, integrating the first 20 genes adjacent to the IR. In the genus *Cabomba*, a similar expansion of the IR_B into the LSC is inferred when compared to *Amborella* or *Nymphaea*, whereby the first 9 genes adjacent to the IR_B were integrated. The plastid genome of *Nuphar advena* is the only genome under study that displays an expansion of the IR_A into the LSC compared to *Amborella* or *Nymphaea*, with the transfer RNA gene for histidine integrated into the IR. The IR regions of the two


Amborella trichopoda
Main Nymphaeaceae* + Brasenia schreberi
Nuphar advena
Cabomba caroliniana
Trithuria inconspicua
Trithuria filamentosa
Illicium oligandrum Schisandra chinensis
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
LSC
IRb
SSC
IRa
photosystem I
photosystem II
cytochrome b/f complex
ATP synthase
RubisCO large subunit
RNA polymerase
ribosomal proteins (SSU)
ribosomal proteins (LSU)
NADH dehydrogenase
clpP, matK
other genes
ycf
ORFs
transfer RNAs
ribosomal RNAs

◀**Fig. 1** Comparison of genome structure and inverted repeat (IR) length among early-diverging angiosperms visualized via aligned linear genome maps. Genes are shown proportional to length and in direction of transcription (*right*: forward, *left*: reverse). Gene *color* indicates functional classification. IR regions are indicated by *dark gray shading*, and large single-copy region (LSC) and small single-copy region (SSC) are indicated by *light gray shading*. Species whose plastomes differ only marginally in length are represented by the same gene map: *Nymphaea ampla*, *N. jamesoniana*, *N. mexicana*, *Victoria cruziana*, *Brasenia schreberi* and *Barclaya longifolia* are represented by *N. alba*; *Illicium oligandrum* is represented by *Schisandra chinensis*

species of the Austrobaileyales under study, by contrast, display a contraction compared to the IRs of *Amborella*, as the IR regions of *Illicium oligandrum* and *Schisandra chinensis* no longer contain the first five genes of the IR_B adjacent to the LSC.

Re-evaluation and correction of annotations

A re-evaluation of gene annotations of the plastid genomes under study was found to be very important toward a correct genomic characterization. Multiple potential annotation errors were identified through this curation procedure. For example, the GenBank record of the plastid genome of *Trithuria inconspicua* (accession NC_020372) appears to confuse the two transfer RNA genes for glycine (i.e., *trn*G-*TCC* and *trn*G-*GCC*) and lacks an annotation for gene *clp*P despite the presence of a translation initiation and a translation termination codon in its location (Table 2). Moreover, this record exhibits IR regions of unequal sequence and length and displays unequal annotations for the two copies of gene *ndh*I, which is duplicated as part of the IR. Similarly, the plastid genome of *Nymphaea* cf. *odorata* appears to confuse the two transfer RNA genes for glycine and displays an incomplete annotation for the transfer RNA gene for cysteine (*trn*C-*GCA*), which is short by approximately 10 nucleotides compared to *Amborella* (Table 2). Moreover, this record has unexpected duplicates of the transfer RNA genes for methionine (*trn*M-*CAT*), threonine (*trn*T-*GGT*) and proline (*trn*P-*GGG*), which are nested within other transfer RNA genes. Similar inconsistencies exist in the genome annotations of other GenBank records analyzed here (Table 2). In addition to these sequence-specific annotation uncertainties, most GenBank records of the 15 plastid genomes analyzed here also lack annotations for the open reading frames *orf*42 and *orf*56, which are located in the intron of the transfer RNA for alanine (*trn*A-*TGC*). Moreover, the start and stop codon positions of several gene annotations in previously sequenced plastid genomes of early-diverging angiosperms appear to be incorrect (Table 2). For example, the following start or stop codon positions appear to be incorrect in *Nymphaea* cf. *odorata* compared to the annotations of the plastid genomes of all other early-diverging angiosperms analyzed, as the latter share the same start and stop codons for the genes in question: Stop codon of exon 1 of gene *atp*F appears to be 12 bp (i.e., 4 amino acids) too late; start codon of exon 2 of gene *pet*B appears to be 15 bp (i.e., 5 amino acids) too early; start codon of exon 2 of gene *pet*D appears to be 51 bp (i.e., 17 amino acids) too early; start codon of exon 2 of gene *rpo*C1 appear to be presumably 9 bp (i.e., 3 amino acids) too early. Similar occurrences of potentially incorrect start and stop codon positions exist for the plastid genomes of *Trithuria filamentosa* and *Trithuria inconspicua* (Table 2).

Gene complement of early-diverging angiosperms

The number of genes in the plastid genomes of early-diverging angiosperms appears to be highly conserved. All input taxa display a set of 116 unique genes in the plastid genome, not counting *orf*188 and the hypothetical protein-coding gene *ycf*68, given that it contains multiple stop codons (Online Resource 6). Of these unique genes, between 14 and 47 are duplicated in the IRs. Among this gene complement, 82 are protein-coding genes, 30 are transfer RNAs (tRNAs), and 8 are ribosomal RNAs (rRNA). A detailed list of genes detected in the plastid genomes under study is given in Table 3, which likely represents the plastid-encoded gene complement of early-diverging angiosperms.

The presence of specific gene features in the plastid genomes of early-diverging angiosperms also appears to be highly conserved. A trans-spliced version of the gene *rps*12, for example, is found among all plastid genomes under study, whereby the gene displays three distinct exons, with the first exon located in LSC, while the second and third exons are duplicated in the IR regions. We detected this exact configuration in all plastid genomes under study except for *Trithuria filamentosa*, where all three exons of the *rps*12 are part of the IR and thus duplicated. Similar to the trans-spliced version of *rps*12, the presence of two open reading frames in the intron of the tRNA for alanine (*trn*A-*TGC*) is detected among all plastid genomes under study. This occurrence has been previously reported by Chumley et al. (2006), who termed these reading frames *orf*42 and *orf*56, respectively. Since *trn*A-*TGC* is located in the IR, both reading frames are duplicated in the plastome. Reading frame *orf*42 displays intact translation initiation and translation stop codons and is devoid of any internal stop codons (Online Resource 7). Reading frame *orf*56 displays one of two possible translation initiation codons, ends with a translation stop codon and exhibits no internal stop codons (Online Resource 8). All taxa under study hereby exhibit the default start codon ATG for *orf*56, except in the plastomes of the genera

Table 1 Comparison of the plastid genomes of early-diverging angiosperms

Plant order	Amborellales	Nymphaeales						
Plant family	Amborellaceae	Nymphaeaceae						
Name of organism	*Amborella trichopoda*	*Barclaya longifolia*	*Nuphar advena*	*Nymphaea alba*	*Nymphaea alba*	*Nymphaea ampla*	*Nymphaea jamesoniana*	*Nymphaea mexicana*
GenBank accession number	NC_005086	KY284156	NC_008788	KU234277	NC_006050	KU189255	NC_031826	NC_024542
Genome size (bp)	162,686	158,360	160,866	159,929	159,930	159,879	158,830	159,962
LSC length (bp)	90,970	88,717	90,379	90,013	90,014	89,951	89,250	90,019
SSC length (bp)	18,414	19,183	18,817	19,562	19,562	19,574	19,360	19,533
IR length (bp)	26,651	25,230	25,835	25,177	25,177	25,177	25,110	25,205
Number of genes	116	116	116	116	116	116	116	116
Number of protein-coding genes (duplicated in IR)	82 (8)	82 (9)	82 (9)	82 (9)	82 (9)	82 (9)	82 (9)	82 (9)
Number of tRNA genes (duplicated in IR)	30 (7)	30 (7)	30 (8)	30 (7)	30 (7)	30 (7)	30 (7)	30 (7)
Number of rRNA genes (duplicated in IR)	4 (4)	4 (4)	4 (4)	4 (4)	4 (4)	4 (4)	4 (4)	4 (4)
Proportion of coding to non-coding regions	0.69	0.69	0.7	0.65	0.7	0.71	0.68	0.69
Average gene density (genes/kb)	0.82	0.85	0.84	0.84	0.84	0.84	0.84	0.84
GC content (%)	38.3	39.1	39.1	39.2	39.2	39.2	39.2	39.1

Plant order	Nymphaeales					Austrobaileyales	
Plant family	Cabombaceae			Hydatellaceae		Schisandraceae	
Name of organism	*Victoria cruziana*	*Brasenia schreberi*	*Cabomba caroliniana*	*Trithuria filamentosa*	*Trithuria inconspicua*[a]	*Illicium oligandrum*	*Schisandra chinensis*
GenBank accession number	KY001813	NC_031343	KT705317	KF696682	NC_020372	NC_009600	KU362793
Genome size (bp)	158,993	158,951	164,057	180,562	165,389	148,553	147,772
LSC length (bp)	89,572	88,757	82,090	68,776	84,468	98,057	97,351
SSC length (bp)	19,535	19,514	18,827	6356	6354	20,268	20,305
IR length (bp)	24,943	25,340	31,570	52,715	37,291	15,114	15,058
Number of genes	116	116	116	116	116	115	115
Number of protein-coding genes (duplicated in IR)	82 (9)	82 (9)	82 (19)	82 (35)	82 (17)	81 (6)	81 (5)
Number of tRNA genes (duplicated in IR)	30 (7)	30 (7)	30 (7)	30 (7)	30 (5)	30 (5)	30 (5)
Number of rRNA genes (duplicated in IR)	4 (4)	4 (4)	4 (4)	4 (4)	4 (4)	4 (4)	4 (4)
Proportion of coding to non-coding regions	0.69	0.69	0.69	0.73	0.72	0.69	0.69
Average gene density (genes/kb)	0.84	0.84	0.88	0.89	0.85	0.88	0.87

Table 1 continued

Plant order	Nymphaeales					Austrobaileyales	
Plant family	Cabombaceae			Hydatellaceae		Schisandraceae	
Name of organism	*Victoria cruziana*	*Brasenia schreberi*	*Cabomba caroliniana*	*Trithuria filamentosa*	*Trithuria inconspicua*[a]	*Illicium oligandrum*	*Schisandra chinensis*
GC content (%)	39.1	39	38.3	39.5	39.6	39	39.5

The potential annotation mistakes listed in Table 3 have been accounted for prior to generating this table. Given our discussion in the main text, the table does not count the hypothetical gene *ycf*68 (annotated in accessions NC_024542 and NC_020372). It also disregards the open reading frame *orf*188 (annotated in accessions NC_005086, NC_020372, NC_009600, KU362793). It does, however, count the hypothetical gene *ycf*15, which is present in all plastid genomes under study except those of the order Austrobaileyales, even though it was not annotated in all genomes. This table also counts the open reading frames *orf*42 and *orf*56

[a] Lengths of the two inverted repeats in *Trithuria inconspicua* differ, and only the longer one is given. Inverted repeats and single-copy regions of the genome thus do not sum up to total genome length

Barclaya, *Nymphaea* and *Victoria*, which display the start codon GTG.

Presence of hypothetical genes *ycf*15 and *ycf*68

Next to the regular set of genes, we also detected two hypothetical protein-coding genes in the plastid genomes under study which are not consistently annotated and described in angiosperm plastomes: *ycf*15 and *ycf*68. Hypothetical gene *ycf*15 is found in all plastid genomes under study except those of the order Austrobaileyales. All analyzed plastid genomes of the Nymphaeales contain the *ycf*15 motif. Compared to the *ycf*15 region of *Nicotiana tabacum*, where it was described for the first time, these copies contain an intervening sequence, possibly an intron, between the flanking regions that may be transcribed (Online Resource 5). The 5′ flanking region of *ycf*15 starts with a GTG start codon, although an alternative ATG start codon exists at position 52–54 of the alignment. All analyzed species contain stop codons earlier than *Nicotiana*. Only *Amborella trichopoda*, *Nymphaea alba* and *Nymphaea* cf. *odorata* are without a stop codon in the first exon and thus display a potential reading frame of 61, 81 and 80 amino acids, respectively, compared to 88 amino acids in *Nicotiana*, assuming that the intervening sequence represents an intron. Both *Trithuria* species have insertions and one deletion in the first exon, leading to reading frame shifts. Similarly, all species except for *Amborella* have gaps in the second exon, leading to reading frame shifts and diverging amino acid sequences. Hypothetical gene *ycf*68 could be detected in all plastid genomes under study. Most sequences display gaps that result in reading frame shifts and internal stop codons. An alternative start codon at position 189 of the alignment results in a relatively long reading frame of 306 bp (i.e., 102 amino acids) in *Nymphaea alba*, *Nymphaea* cf. *odorata*, *Nuphar advena* and *Brasenia schreberi*, which may be translated (Online Resource 6).

Phylogenetic analysis and data partitioning

Our phylogenetic reconstructions of the multi-gene alignment resulted in highly resolved phylogenetic trees, with almost all clades recovered having full branch support (Fig. 2). Monophyly of the families Cabombaceae and Hydatellaceae was fully supported in each inference (BS 100, PP 1.00), whereas monophyly of the family Nymphaeaceae was not. The position of the genus *Nuphar* in relation to the rest of the Nymphaeaceae and the Cabombaceae was not resolved. All partitioning schemes resulted in very weak support for the position of *Nuphar*, with three of them indicating a clade of *Nuphar* and the Cabombaceae (Online Resource 9): BS 58/PP 0.85 in the unpartitioned

Table 2 Overview of potential errors and instances for improvement in gene annotations of several previously published plastid genomes of early-diverging angiosperms

Species name (GenBank accession number)	Annotation tool	Potential errors and instances for improvement in gene annotations
Amborella trichopoda (NC_005086)	n.s.	The hypothetical gene in position 33,808–33,701 likely constitutes gene *psb*M
		The hypothetical gene in position 65,359–65,469 likely constitutes gene *psa*I
		The hypothetical gene in position 72,508–72,603 likely constitutes gene *pet*L
		The annotation of tRNA *trn*I-*CAT* appears to be missing
Nymphaea mexicana (NC_024542)	DOGMA	Separate annotations of tRNA *trn*M-*CAT* exist, of which one (34,719–34,777) occurs at the position of *trn*T-*GGT* and appears to be invalid
		Separate annotations of tRNA *trn*T-*GGT* exist, of which one (56,975–57,033) occurs at the position of *trn*M-*CAT* and appears to be invalid
		The annotation of tRNA *trn*P-*GGG* exists twice at the same position
		The annotation of tRNA *trn*G-*GCC* appears to be incorrectly labeled as *trn*G-*TCC*
		The annotation of tRNA *trn*C-*GCA* appears to end prematurely by approx. 10 bp compared to *Amborella*
		Compared to the annotations of the other 14 plastid genomes analyzed, the following start or stop codon positions appear to be incorrect in *Nymphaea mexicana*:
		Stop codon of exon 1 of gene *atp*F appears to be 12 bp (i.e., 4 amino acids) too far downstream
		Start codon of exon 2 of gene *pet*B appears to be 15 bp (i.e., 5 amino acids) too far upstream
		Start codon of exon 2 of gene *pet*D appears to be 51 bp (i.e., 17 amino acids) too far upstream
		Start codon of exon 2 of gene *rpo*C1 appears to be 9 bp (i.e., 3 amino acids) too far upstream
Trithuria filamentosa (KF696682)	DOGMA	Compared to the annotations of the other 14 plastid genomes analyzed, the following start or stop codon positions appear to be incorrect in *Trithuria filamentosa*:
		Start codon of gene *acc*D appears to be 57 bp (i.e., 19 amino acids) too far upstream
		Start codon of gene *psb*C appears to be 36 bp (i.e., 12 amino acids) too far upstream, which, when corrected, would have the typical start codon (ATG) instead of the current atypical one (GTG)
Trithuria inconspicua (NC_020372)	n.s.	The annotation of gene *clp*P appears to be missing
		The annotation of tRNA *trn*G-*GCC* appears to be incorrectly labeled as *trn*G-*TCC* and vice versa
		The annotation of tRNA *trnf*M-*CAT* appears to be incorrectly labeled as *trn*M-*CAT*
		The annotation name of tRNA *trn*I-*CAT* appears to require correction
		The annotations of gene *ndh*I, which is located in the IR, are of unequal length
		The annotation of the $5'$ end of gene *ndh*D is different in the IR$_{\text{A}}$ than in the IR$_{\text{B}}$
		Compared to the annotations of the other 14 plastid genomes analyzed, the following start or stop codon positions appear to be incorrect in *Trithuria inconspicua*:
		Start codon of gene *psb*H appears to be 18 bp (i.e., 6 amino acids) too far upstream.
		Start codon of gene *ndh*H appears to be 24 bp (i.e., 8 amino acids) too far upstream.
		Start codon of gene *ndh*I appears to be 15 bp (i.e., 5 amino acids) too far upstream.
		Start codon of gene *ycf*2 appears to be 159 bp (i.e., 53 amino acids) too far upstream.
Illicium oligandrum (NC_009600)	DOGMA	The annotation of tRNA *trn*P-*TGG* appears to be incorrectly labeled as *trn*P-*GGG*
Schisandra chinensis (KU362793)	Unpublished record	The annotation of tRNA *trn*P-*TGG* appears to be incorrectly labeled as *trn*P-*GGG*

The annotation tools employed are listed as specified in the publication that is associated with the GenBank record

approx., approximately; n.s., not specified

matrix, BS 65/PP 0.88 in the matrix of one partition per codon and BS 55/PP 0.97 in the matrix inferred as optimal via PartitionFinder2. The ML tree of the partitioned-by-gene matrix displayed *Nuphar* as sister to a clade formed by the Cabombaceae and the rest of the Nymphaeaceae, but without statistical support (BS 33, PP 0.22). The phylograms that resulted under the different data partitioning strategies displayed little divergence in the lineage between the stem lineages of *Nuphar* and the Cabombaceae (Online Resource 9). Furthermore, *Victoria cruziana* was found

Table 3 Presumptive gene complement of early-diverging angiosperms

Functional class	Number of genes	Gene names	Protein subunits
Ribosomal RNAs	4	*rrn*4.5(2x), *rrn*5(2x), *rrn*16(2x), *rrn*23(2x)	4.5S, 5S, 16S, 23S
Transfer RNAs	30	tRNA-Ala$^{(TGC)}$(2x)*, tRNA-Asn$^{(GTT)}$(2x), tRNA-Arg$^{(ACG)}$(2x), tRNA-Arg$^{(TCT)}$, tRNA-Asp$^{(GTC)}$, tRNA-Cys$^{(GCA)}$, tRNA-fMet$^{(CAT)}$, tRNA-Gln$^{(TTG)}$, tRNA-Glu$^{(TTC)}$, tRNA-Gly$^{(GCC)}$, tRNA-Gly$^{(TCC)}$, tRNA-His$^{(GTG)}$[a], tRNA-Ile$^{(CAT)}$[b], tRNA-Ile$^{(GAT)}$(2x)*, tRNA-Leu$^{(CAA)}$[b], tRNA-Leu$^{(TAA)}$, tRNA-Leu$^{(TAG)}$, tRNA-Lys$^{(TTT)}$*, tRNA-Met$^{(CAT)}$, tRNA-Phe$^{(GAA)}$, tRNA-Pro$^{(TGG)}$, tRNA-Ser$^{(GCT)}$, tRNA-Ser$^{(GGA)}$, tRNA-Ser$^{(TGA)}$, tRNA-Thr$^{(GGT)}$, tRNA-Thr$^{(TGT)}$, tRNA-Trp$^{(CCA)}$, tRNA-Tyr$^{(GTA)}$, tRNA-Val$^{(GAC)}$(2x), tRNA-Val$^{(TAC)}$	
Photosystem I	5	*psa*A, *psa*B, *psa*C[c], *psa*I, *psa*J	A, B, C, I, J
Photosystem II	15	*psb*A, *psb*B[d], *psb*C, *psb*D, *psb*E, *psb*F, *psb*H[d], *psb*I, *psb*J, *psb*K, *psb*L, *psb*M, *psb*N[d], *psb*T[d], *psb*Z	A, B, C, D, E, F, H, I, J, K, L, M, N, T, Z
Cytochrome b6/f complex	6	*pet*A, *pet*B[d], *pet*D[d], *pet*G, *pet*L, *pet*N	A, B, D, G, L, N,
ATP synthase	6	*atp*A, *atp*B, *atp*E, *atp*F, *atp*H, *atp*I	A, B, E, F, H, I
Ribulose-1,5-bisphosphate carboxylase/oxygenase	1	*rbc*L	L
NADH dehydrogenase	11	*ndh*A[c], *ndh*B(2x), *ndh*C, *ndh*D[c], *ndh*E[c], *ndh*F[c], *ndh*G[c], *ndh*H[c], *ndh*I[c], *ndh*J, *ndh*K	A, B, C, D, E, F, G, H, I, J, K
Ribosomal protein (large subunit)	9	*rpl*2[b], *rpl*14[d,e], *rpl*16[d,e], *rpl*20, *rpl*22[d,e], *rpl*23[b], *rpl*33, *rpl*32[c], *rpl*36[d,e]	2, 14, 16, 20, 22, 23, 33, 32, 36
Ribosomal protein (small subunit)	12	*rps*2, *rps*3[d,e], *rps*4, *rps*7(2x), *rps*8[d,e], *rps*11[d,e], *rps*12(2x)*[d], *rps*14, *rps*15[c], *rps*16, *rps*18, *rps*19[d,e]	2, 3, 4, 7, 8, 11, 12, 14, 15, 16, 18, 19
RNA polymerase	4	*rpo*A[d,e], *rpo*B, *rpo*C1, *rpo*C2	A, B, C1, C2
ATP-dependent protease	1	*clp*P	P
Cytochrome c biogenesis	1	*ccs*A[c]	A
Membrane protein	1	*cem*A	A
Maturase for group II introns	1	*mat*K	K
Initiation factor	1	*inf*A[d,e]	A
Acetyl-CoA carboxylase	1	*acc*D	D
Hypothetical protein-coding gene		*ycf*1[c], *ycf*2[b], *ycf*3, *ycf*4, *ycf*15[f]	
Open reading frames		*orf*42(2x), *orf*56(2x)	

This list indicates all the genes encoded in the 15 plastid genomes of the early-diverging angiosperms under study. Genes indicated by "2x" are located in the inverted repeats in all taxa under study and are thus duplicated. Genes indicated by an asterisk contain one or more introns. As in Table 2, this table does not count the hypothetical gene *ycf*68 nor the open reading frame *orf*188

[a] Duplicated in *Nuphar advena*

[b] Duplicated in all early-diverging angiosperms under study except in the members of the Austrobaileyales

[c] Duplicated only in the members of the Hydatellaceae under study

[d] Duplicated in *Trithuria filamentosa*

[e] Duplicated in *Cabomba caroliniana*

[f] Missing in the members of the Hydatellaceae under study

nested within the genus *Nymphaea* in each phylogenetic reconstruction and with high statistical support, indicating paraphyly of the genus *Nymphaea* in its current circumscription. All phylograms also indicated a long stem lineage of *Trithuria* compared to all other plastid genomes analyzed.

Taxon designation of GenBank record NC_024542

Our reconstruction of the phylogenetic relationships between the 15 taxa under study, the 71 accessions of *Nymphaea* and the 16 accessions of related genera to evaluate the taxon designation of GenBank record

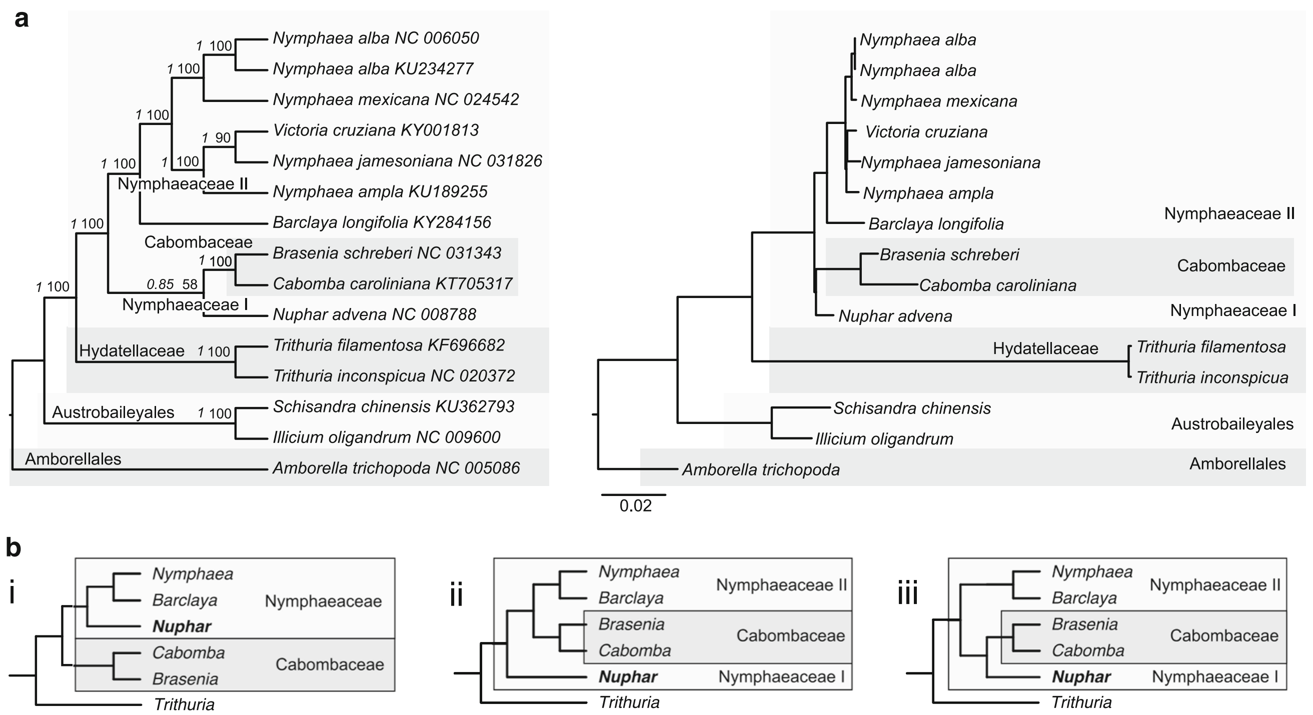


Fig. 2 Inferred phylogenetic relationships of 15 early-diverging angiosperm taxa and the potential phylogenetic position of the genus *Nuphar*. **a** Phylogenetic relationships of the 15 taxa under study as inferred from a multi-gene dataset of 77 plastid-encoded genes under ML phylogeny inference without data partitioning, visualized as cladogram (left) and as phylogram (right). Bootstrap support values of the ML inference greater than 50% and posterior probabilities (*in italics*) of a concomitant BI phylogeny inference greater than 0.5 are given above branches of the cladogram. **b** Three potential scenarios for the phylogenetic position of *Nuphar* within the Nymphaeales based on the results of our phylogenomic analyses that included data partitioning

NC_024542 (Yang et al. 2014) using *trn*T–*trn*F as phylogenetic marker indicated that the plastid genome in question unlikely represents *Nymphaea mexicana*. The sequence in question is recovered as sister to a specimen of *Nymphaea odorata* subsp. *tuberosa* (BS 78, PP 0.99) and embedded in a clade comprised exclusively of specimens of *N. odorata* (Fig. 3). The statistical support for this clade is medium to high (BS 65, PP 0.98). Our reconstructions also infer a sister relationship between the species *Nymphaea odorata* and *N. mexicana*, which is supported by maximum branch support (BS 100, PP 1.00). The phylogeny further recovers a fully supported clade of Cabombaceae and Nymphaeaceae pro parte without *Nuphar* (BS 100, PP 1.00).

Discussion

Gene content and synteny in plastomes of early-diverging angiosperms

The gene content of land plants is relative stable, with gene losses in the angiosperms mainly associated with parasitic or heteromycotrophic lifestyles (Logacheva et al. 2011; Wicke et al. 2011; Cusimano and Wicke 2016). All taxa included in this investigation are photoautotrophic, and their gene content is almost identical. *Amborella* and all species of the Nymphaeales analyzed exhibit 116 genes (not including *ycf*68 and *orf*188), whereas the two species of Austrobaileyales lack the hypothetical gene *ycf*15 and thus exhibit 115 genes (Table 1). Major structural changes have been described from several angiosperm lineages (Haberle et al. 2008; Cai et al. 2008; Weng et al. 2014), but such changes were not found in the plastid genomes of the Nymphaeales.

Generally, the sizes of plastid genomes can vary considerably, from 63 kbp in the parasitic *Phelipanche* to 242 kbp in the highly rearranged plastomes of *Pelargonium* (Weng et al. 2014). More than half of the 1983 plastid genomes of seed plants that are currently available on GenBank (GenBank search on February 11, 2017, for “chloroplast genomes” of Spermatophyta with genome size >60 kbp) display a genome size between 140 and 160 kbp. With sizes between 158,360 and 160,866 bp, the complete lengths of the plastid genomes of the Nymphaeaceae species analyzed are on the upper end of this spectrum. In the Cabombaceae, *Brasenia schreberi* has a plastid genome size similar to members of Nymphaeaceae, whereas *Cabomba caroliniana* displays a larger plastome (164,057 bp). The plastid genome of *Trithuria inconspicua*

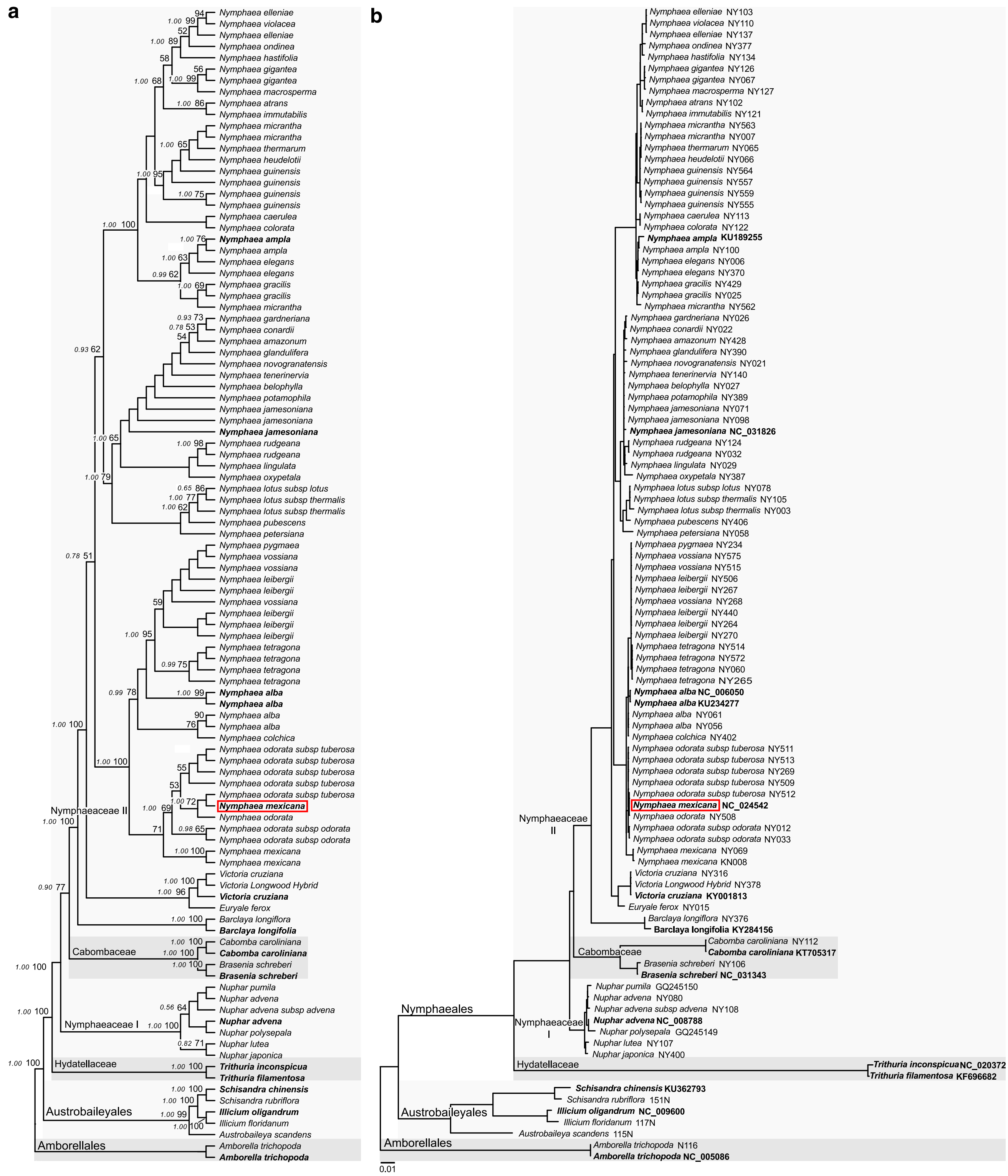


Fig. 3 Phylogenetic position of the 15 plastid genomes analyzed among a set of 71 accessions of *Nymphaea* and 16 accessions of related early-diverging angiosperms as inferred from a DNA sequence alignment of *trn*T–*trn*F. The displayed trees both constitute the tree with the highest likelihood score inferred, displayed as cladogram under (**a**) and as phylogram under (**b**). The sequences from the plastid genomes are highlighted in *bold*, and those from the genome with GenBank accession number NC_024542 ("*Nymphaea mexicana,*" Yang et al. 2014) are additionally highlighted in *red*. Bootstrap support values of the ML inference greater than 50% and posterior probabilities (*in italics*) of a concomitant BI phylogeny inference greater than 0.5 are given above *branches* in **a**. The branch from *Amborella trichopoda* to the rest of the species analyzed is truncated for better visualization in **b**

has an even larger size (165,389 bp), while the plastome of *T. filamentosa* (180,562 bp) exceeds the size of all but ten publicly available plastid genomes. By comparison, the plastomes of both Austrobaileyales under study were found considerably smaller (<150 kbp). While plastome size reductions are common in parasitic plants as a result of gene loss (Wicke et al. 2013), size increases are usually the result of expansions of the IR regions (Weng et al. 2014). Similarly, the differences in plastome size recorded in the present investigation are mainly the result of extractions or contractions of the IR regions. It is noteworthy that the plastid genomes of the two species of *Trithuria* under study display a length difference of >15 kbp, despite being likely sister species. According to Iles et al. (2014), the two species have diverged very recently (0.5 Ma [0–1.1 Ma]). Given this recent time of divergence, it is likely that the IR expansion between these two species originated via a single change in length rather than a gradual increase in IR size.

The gene content and the gene synteny of the plastid genome sequences analyzed here are identical. Even though the topology of early-diverging angiosperm lineages is still disputed (Drew et al. 2014; Goremykin et al. 2015; Simmons 2016), the majority of studies support *Amborella* as sister to all other angiosperms, with Nymphaeales as the second and Austrobaileyales as the third diverging lineage (APW3, Stevens 2017). The conserved gene content and synteny in *Amborella* as well as the Nymphaeales leads to the hypothesis that the most recent common ancestor of both lineages, which may include all extant angiosperms, shared the same gene content and the same gene synteny, with the IRs spanning from *rpl*2 to *ndh*F. Our observed differences in IR length may have originated through length changes on distal branches, such as the IR expansions in the Hydatellaceae and the Cabombaceae or the IR contractions in the Austrobaileyales, but more research is needed to evaluate this hypothesis.

Correction of annotations and SSC orientation in *Trithuria*

The high degree of structural conservation across the plastid genomes of early-diverging angiosperms was initially obscured by ambiguous and probably incorrect annotations in several previously published plastid genomes (Table 2). Our manual corrections of the annotations of these previously published as well as our newly generated plastid genomes were instrumental in obtaining a more correct assessment of gene content and synteny. However, not all ambiguities in gene annotations or genome structure among the previously published genomes are the result of human error. For example, the orientation of the SSC in *Trithuria* appears to be inverted compared to the plastid genomes of other early-diverging angiosperms at first glance, but the relative orientation of the SSC is not fixed in plastid genomes, and both orientations may occur in the same individual (Palmer 1983). Although this form of plastid heteroplasmy has been confirmed in a wide variety of plant species (see Walker et al. 2015 for review), several recent investigations have overlooked this fact and as a result pronounced the SSC a hotspot of inversions (Walker et al. 2015). Consequently, the different orientation of the SSC in *Trithuria* compared to the plastid genomes of other early-diverging angiosperms is unlikely the result of a genomic rearrangement and more likely the observation of natural plastid heteroplasmy.

Open reading frames in early-diverging angiosperms

While the functions of most genes in the plastome are well understood (Wicke et al. 2011), several hypothetical reading frames of unknown function remain. Most plastid genomes contain a varying number of open reading frames (*orf*) that have different degrees of conservation. More conserved *orf* with similar amino acid content are called hypothetical chloroplast open reading frames and hypothetical protein-coding genes (*ycf*).

The presence of *orf*42 and *orf*56 in the plastid genomes of early-diverging angiosperms has been acknowledged by some but not all investigations that generated the previously published plastid genomes. We located the open reading frame *orf*42 in the plastid genomes of all 15 plastid genomes analyzed. Each of them displays intact translation initiation and translation stop codons and no internal stop codons (Online Resource 7). The open reading frame is located inside the intron of the tRNA for alanine (*trn*A-*TGC*) and, according to Chumley et al. (2006), has high sequence similarity with the 3′ end of the gene *pvs-trn*A, which in *Phaseolus* is found in the mitochondrial genome (Woloszynska et al. 2004). Similarly, the open reading frame *orf*56 is also present in the intron of *trn*A-*TGC* in all plastid genomes under study, is reported only from the plastomes of *Nymphaea mexicana* and *Trithuria inconspicua* and does not start with the common translation initiation codon (ATG) in all taxa (Online Resource 8). Instead, *orf*56 appears to start with the codon GTG in the plastomes of *Barclaya*, *Nymphaea* and *Victoria*. The codon GTG constitutes one of two alternative translation initiation codons that have been reported from plastid genes of angiosperms (Kuroda et al. 2007; Raubeson et al. 2007). According to Chumley et al. (2006), *orf*56 displays high sequence similarity with the gene *ACRS*, which is found in the mitochondrial genome in *Citrus* (Ohtani et al. 2002). Both *orf*42 and *orf*56 are found in numerous angiosperm plastid genomes, including *Pelargonium* (Chumley et al. 2006) and *Veratrum* (Do et al. 2013) and could thus be part of the default gene complement of angiosperm plastomes. Goremykin et al. (2003b) described the *orf*56 in *Calycanthus* and also remarked the similarity to the mitochondrial gene *ACRS* from *Citrus*. However, *orf*42 and *orf*56 do not seem to be translated to functional proteins throughout the angiosperms. In plastomes of the genus *Utricularia*, for example, *orf*42 and *orf*56 are reported as pseudogenes due to frameshifts and internal stop codons that are also absent from nuclear or mitochondrial genome contigs (Silva et al. 2016). We do not suggest that *orf*42 or *orf*56 necessarily constitute functional genes in the plastid genomes of early-diverging angiosperms and, at this point, have insufficient information to determine if their transcripts are translated to functional proteins. We do, however, wish to point out that the plastomes of all taxa evaluated in this investigation contain intact copies of both open reading frames, assuming that the codon GTG mediates successful translation initiation for *orf*56. It is noteworthy that those studies that have reported the presence of these hypothetical genes had taken additional steps toward improved gene detection and annotation refinement, such as the correction of gene annotations or the application of supplementary annotation software (e.g., Do et al. 2013; Silva et al. 2016).

The origin of the hypothetical gene *ycf*15 has been addressed by various earlier investigations. It was first described as *orf*87 in *Nicotiana* (Shinozaki et al. 1986) and subsequently examined with regard to its functionality as a protein-coding gene by Schmitz-Linneweber et al. (2001), who found that *ycf*15 occurs in *Spinacia oleracea*, *Arabidopsis thaliana, Zea mays* and *Oenothera berteriana*. The hypothetical gene was hereby found to contain an intervening sequence of 250–300 bp, which is absent in other plants like *Nicotiana tabacum* and was suspected to be an intron (Schmitz-Linneweber et al. 2001). The intervening sequence was hypothesized to be present in early-diverging angiosperms and that it was lost in several lineages throughout angiosperm diversification. Raubeson et al. (2007) compared the occurrence of *ycf*15 in a wider sample of angiosperms and showed that the intervening sequence is indeed present in several early angiosperms including *Amborella*, *Nymphaea*, *Nuphar*, most monocots and some eudicots, but that it was lost in all asterids investigated. The much deeper sampling of Nymphaeales in this study corroborates the conclusions by Schmitz-Linneweber et al. (2001) and Raubeson et al. (2007) for all *Nymphaea* species under study as well as the Cabombaceae and Hydatellaceae. The question of successful translation of *ycf*15 is less clear, however. The sequencing of cDNA by Schmitz-Linneweber et al. (2001) indicated a full translation of the flanking regions of *ycf*15, but internal stop codons would likely interrupt the coding of polypeptides. Large portions of the open reading frame present in *Nicotiana* occur only in *Amborella trichopoda*, *Nymphaea alba* and *N.* cf. *odorata*, while stop codons and frame shifts interrupt the reading frame in all other investigated taxa. The fact that *ycf*15 is partially preserved might reflect functional relevance, for example as a promoter or terminator sequence in gene regulation.

The hypothetical gene *ycf*68 is found in all plastid genomes under study but consistently contains internal stop codons in many angiosperms (Raubeson et al. 2007). In *Utricularia*, for example, the gene *ycf*68 has a frameshift and one in-frame stop codon and may thus only be transcribed but not translated (Silva et al. 2016). Among the taxa studied here, *ycf*68 also displays multiple internal stop codons. Goremykin et al. (2004) proposed a later start for *Nymphaea alba* (at nucleotide position 189), which results in an amino acid sequence without internal stop codons. It is likely that the same applies to the *ycf*68 sequences of other *Nymphaea* species and of *Nuphar*. The idea of Goremykin et al. (2004) should thus be given greater consideration. The sequences of *Barclaya*, *Victoria* and both *Trithuria* species, however, would maintain internal stop codons even under this alternative start position due to indels and subsequent reading frame shifts compared to other Nymphaeaceae plastomes.

Phylogenetic position of *Nuphar* and potential paraphyly of the Nymphaeaceae

Based on a multi-gene alignment of 77 plastid-encoded genes extracted from the 15 plastid genomes under study, we conducted phylogenetic reconstructions under different data partitioning strategies. In each of these reconstructions, the families Cabombaceae and Hydatellaceae are supported as monophyletic, whereas the family Nymphaeaceae is not (Fig. 2). The phylogenetic position of the genus *Nuphar* remained unresolved. Under an unpartitioned matrix, for example, *Nuphar* is recovered as sister to the Cabombaceae, but the respective node is merely supported by 0.85 PP under BI and 58% BS under ML. The extremely short branch subtending the most recent common ancestor of *Nuphar* and the Cabombaceae in this matrix furthermore indicates a rapid diversification (Fig. 2, phylogram); other partitioning strategies produced very similar results (Online Resource 9). Three potential scenarios therefore remain for the phylogenetic position of *Nuphar* within the Nymphaeales: *Nuphar* as an early-diverging lineage of the Nymphaeaceae (Fig. 2b-i); *Nuphar* as sister to a clade formed by the Cabombaceae and the rest of Nymphaeaceae (Fig. 2b-ii); and *Nuphar* forming a clade with the Cabombaceae (Fig. 2b-iii). The latter two scenarios render the family Nymphaeaceae paraphyletic. Thus, phylogenetic analyses are needed that include a more extensive taxon sampling of *Nuphar* and the Cabombaceae at the species level and maybe the additional use of plastid spacers and introns.

Previous investigations had also been unable to consistently resolve the phylogenetic position of *Nuphar*. The monophyly of the Nymphaeaceae including *Nuphar* was supported by a combined dataset of nine plastid regions including introns, intergenic spacers and *mat*K (Löhne et al. 2007). However, an earlier analysis of *trn*T–*trn*F sequences found only limited support for such a clade (Borsch et al. 2003), although it did not include DNA sequences of the Hydatellaceae. A study based on concatenated DNA sequences of *matK* and ITS2 available on GenBank, including accessions of the Hydatellaceae, found support for a clade comprising of *Nuphar* and the Cabombaceae (Biswal et al. 2012). However, the results of Biswal et al. (2012) should be interpreted with care, as their taxon sampling is heavily asymmetric: None of the other early-branching angiosperms (*Amborella*, Austrobaileyales) were included in their dataset except the Nymphaeales, and gymnosperms were used as outgroup.

From a morphological perspective, a scenario in which the Nymphaeaceae are monophyletic and sister to Cabombaceae currently remains the most plausible solution. A parsimony analysis of 66 morphological characters, including data on the Hydatellaceae, placed *Nuphar* as sister to the remainder of Nymphaeaceae (Borsch et al. 2008). Furthermore, eusyncarpous carpels were hypothesized to have arisen in the common ancestor of the Nymphaeaceae, while the Cabombaceae and the Hydatellaceae have apocarpous carpels (Borsch et al. 2008). Eusyncarpy also occurs in *Illicum* and the eudicots (Doyle and Endress 2000; Rudall et al. 2007), suggesting multiple gains of this feature in angiosperms. The alternative hypothesis of a sister relationship between *Nuphar* and the Cabombaceae (and also the third topology with *Nuphar* and Cabombaceae as successive sisters to the core Nymphaeaceae) would require a more complex explanation for the evolution of the gynoecium in the Nymphaeales. In the event of a sister relationship between *Nuphar* and the Cabombaceae, a less than parsimonious solution would also apply to pollen evolution, as the granular-intermediate infratectum is considered a synapomorphy of all Nymphaeaceae (Borsch et al. 2008).

As indicated by long branches of the presented phylogenies (Figs. 2, 3), the two representatives of the Hydatellaceae display numerous autapomorphic nucleotide changes compared to the other taxa under study. In order to determine if the phylogenetic position of *Nuphar* was recovered as the earliest-diverging genus of the Nymphaeaceae and Cabombaceae (Figs. 2, 3) represents an artifact of long-branch attraction to the Hydatellaceae, we repeated our phylogeny inference on the *trn*T–*trn*F sequence alignment after excluding the DNA sequences of the Hydatellaceae. The resulting phylogeny did not deviate from the initial one, indicating that *trn*T–*trn*F either contains a different phylogenetic signal than the multi-gene alignment of 77 plastid-encoded genes or a different level of homoplasy.

The results of our investigation support the hypothesis that the genus *Nymphaea* is paraphyletic in its current circumscription. A clade comprised of *Barclaya*, *Nymphaea*, *Victoria* and *Euryale* was supported by Borsch et al. (2007) and by Löhne et al. (2008). The shift in translation initiation codon of the open reading frame *orf*56 from ATG to GTG may be a molecular synapomorphy for this clade (Online Resource 8). While the first well-sampled molecular phylogenies of the Nymphaeales based on *trn*T–*trn*F (Borsch et al. 2007) inferred the *Victoria-Euryale* clade as sister to a weakly supported monophyletic genus *Nymphaea* (including *Ondinea*, see Löhne et al. 2009), the addition of further plastid sequence data (Löhne et al. 2007) recovered it as sister to a clade comprised of *Nymphaea* subg. *Hydrocallis* and *Nymphaea* subg. *Lotus,* although with only moderate branch support. The phylogenetic trees inferred from our alignment of 77 plastid-encoded genes strongly support the placement of *Victoria* within *Nymphaea* and also support the first branching position of the temperate subclade of *Nymphaea* (Fig. 2). Thus, it appears that genome-scale data hold great promise to further illuminate the relationships and evolutionary diversification of the water lilies.

Taxon designation of GenBank record NC_024542

Doubts about the correct identification of a previously published plastid genome of a species of *Nymphaea* arose upon the comparison of standard phylogenetic DNA markers across this and other *Nymphaea* samples. The GenBank record in question was published by Yang et al. (2014) as part of an investigation on primer development for the long-range PCR amplification of plastid genomes and was designated as "*Nymphaea mexicana.*" Our phylogenetic tree based on *trn*T–*trn*F recovered this sequence as nested within the North American clade of *Nymphaea odorata* and as sister to a specimen from Vermont (USA), which belongs to *Nymphaea odorata* subsp. *tuberosa*. The *trn*T–*trn*F sequence of NC_024542 is completely identical to the sequence of the Vermont specimen, even in the hypervariable parts of the P8 stem loop (data not shown), which were excluded from the alignment. Unfortunately, we were unable to view and assess the herbarium record associated with NC_024542, as the herbarium voucher could not be located by the original institution (E. Liu, pers. comm.). Although both species (*N. mexicana* and *N. odorata*) are monophyletic (Borsch et al. 2014), there are hybrids between them, and plants of *Nymphaea odorata* have occasionally been crossed into other species to breed ornamental specimens. Some of these ornamental plants have a yellowish flower color, and it seems possible that such an ornamental individual may have been used for generating the plastid genome NC_024542. The uncertainty around the taxonomic identity of this plastid genome illustrates that the careful identification of plant material and the generation of publicly available herbarium specimens for taxonomic re-evaluation remain important tasks in the process of phylogenomic analysis.

Conclusion

The plastid genomes of early-diverging angiosperms were found to be highly conserved with regard to gene content and gene synteny. The full degree of conservation did, however, only become apparent after the manual correction of the annotations of several previously published plastid genomes under study, underscoring the need for the re-assessment of genome annotations that have not undergone stringent manual curation. Our phylogenetic reconstructions revealed a potentially paraphyletic family Nymphaeaceae and a paraphyletic genus *Nymphaea*, which stand in contrast to the results of prior molecular and morphological studies and indicate the need for further investigation. In particular, the phylogenetic position of the genus *Nuphar* requires additional assessment. Methodologically, our results indicate that plastid genomics in the Nymphaeales offers great potential for further insights into the evolutionary diversification of the water lily clade. In particular, the fully conserved gene order has the potential to provide a study case for genome-level alignments that include non-coding regions.

Acknowledgements The authors would like to thank Virginia Duwe of the Botanischer Garten und Botanisches Museum Berlin and Bettina Giesicke of the Freie Universität Berlin for assistance with DNA extractions. The authors also thank Susan Mbedi of the Berlin Center for Genomics in Biodiversity Research for assistance with Illumina library preparations. Moreover, the authors would like to thank the gardeners of the Botanischer Garten und Botanisches Museum Berlin for cultivating the water lily living collection in the greenhouses that provided the specimens of the taxa under study. Finally, the authors would like to acknowledge the high-performance computing service of the ZEDAT of the Freie Universität Berlin for providing allocations of computing time. This investigation was supported by a start-up grant (Initiativmittel der Forschungskommission) of the Freie Universität Berlin to MG.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Information on Electronic Supplementary Material

Online Resource 1. Species name, taxonomic position, place of publication, GenBank accession number and other associated information for each of the 15 plastid genomes analyzed.
Online Resource 2. Overview of read number, mean read length, coverage depth, contig number, contig length and other assembly statistics of plastid genomes that were newly-generated for this investigation.
Online Resource 3. Circular plastome maps of *Nymphaea alba* (KU234277), *Barclaya longifolia* (KY284156), *Nymphaea ampla* (KU189255), *Nymphaea jamesoniana* (NC_031826), *Victoria cruziana* (KY001813), *Brasenia schreberi* (NC_031343) and *Cabomba caroliniana* (KT705317).
Online Resource 4. Individual partitions and their best-fitting nucleotide substitution models of the four data partitioning strategies.
Online Resource 5. Alignment of the hypothetical gene *ycf*15 across the study taxa.
Online Resource 6. Alignment of hypothetical gene *ycf*68 across the study taxa.
Online Resource 7. Alignment of the open reading frame *orf*42 across the study taxa.
Online Resource 8. Alignment of the open reading frame *orf*56 across the study taxa.
Online Resource 9. Best ML trees inferred from the multi-gene dataset of 77 plastid-encoded genes under four different data partitioning strategies.

References

Biswal DK, Debnath M, Kumar S, Tandon P (2012) Phylogenetic reconstruction in the order Nymphaeales: ITS2 secondary structure analysis and in silico testing of maturase k (*matK*) as a potential marker for DNA bar coding. BMC Bioinform 13(Suppl 17):S26. doi:10.1186/1471-2105-13-S17-S26

Borsch T, Hilu KW, Quandt D, Wilde V, Neinhuis C, Barthlott W (2003) Non-coding plastid *trnT–trnF* sequences reveal a well resolved phylogeny of basal angiosperms. J Evol Biol 16:558–576. doi:10.1046/j.1420-9101.2003.00577.x

Borsch T, Hilu KW, Wiersema JH, Löhne C, Barthlott W, Wilde V (2007) Phylogeny of *Nymphaea* (*Nymphaeaceae*): evidence from substitutions and microstructural changes in the chloroplast *trnT–trnF* region. Int J Pl Sci 168:639–671. doi:10.1086/513476

Borsch T, Löhne C, Wiersema J (2008) Phylogeny and evolutionary patterns in Nymphaeales: integrating genes, genomes and morphology. Taxon 57:1052–1081

Borsch T, Löhne C, Mbaye MS, Wiersema J (2011) Towards a complete species tree of *Nymphaea*: shedd, its relationships to the Australian waterlilies. Telopea 13:193–217

Borsch T, Wiersema JW, Hellquist CB, Löhne C, Govers K (2014) Speciation in North American water lilies: evidence for the hybrid origin of the newly discovered Canadian endemic *Nymphaea loriana* sp. nov. (Nymphaeaceae) in a past contact zone. Botany 92:867–882. doi:10.1139/cjb-2014-0060

Cai Z, Guisinger M, Kim H-G, Ruck E, Blazier JC, McMurtry V, Kuehl JV, Boore J, Jansen RK (2008) Extensive reorganization of the plastid genome of *Trifolium subterraneum* (Fabaceae) is associated with numerous repeated sequences and novel DNA insertions. J Molec Evol 67:696–704. doi:10.1007/s00239-008-9180-7

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinform 10:421. doi:10.1186/1471-2105-10-421

Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ, Boore JL, Jansen RK (2006) The complete chloroplast genome sequence of *Pelargonium* × *hortorum*: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. Molec Biol Evol 23:2175–2190. doi:10.1093/molbev/msl089

Cusimano N, Wicke S (2016) Massive intracellular gene transfer during plastid genome reduction in nongreen Orobanchaceae. New Phytol 210:680–693. doi:10.1111/nph.13784

Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. Nature Meth 9:772. doi:10.1038/nmeth.2109

Do HDK, Kim JS, Kim J-H (2013) Comparative genomics of four Liliales families inferred from the complete chloroplast genome sequence of *Veratrum patulum* O. Loes. (Melanthiaceae). Gene 530:229–235. doi:10.1016/j.gene.2013.07.100

Doyle JA, Endress PK (2000) Morphological phylogenetic analysis of basal angiosperms: comparison and combination with molecular data. Int J Pl Sci 161:S121–S153. doi:10.1086/317578

Drew BT, Ruhfel BR, Smith SA, Moore MJ, Briggs BG, Gitzendanner MA, Soltis PS, Soltis DE (2014) Another look at the root of the angiosperms reveals a familiar tale. Syst Biol 63:368–382. doi:10.1093/sysbio/syt108

Gordon A (2014) FASTX-Toolkit. Available at: http://hannonlab.cshl.edu/fastx_toolkit/

Goremykin VV, Hirsch-Ernst KI, Wolfl S, Hellwig FH (2003a) Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. Molec Biol Evol 20:1499–1505. doi:10.1093/molbev/msg159

Goremykin V, Hirsch-Ernst KI, Wölfl S, Hellwig FH (2003b) The chloroplast genome of the "basal" angiosperm *Calycanthus fertilis*—structural and phylogenetic analyses. Pl Syst Evol 242:119–135. doi:10.1007/s00606-003-0056-4

Goremykin VV, Hirsch-Ernst KI, Wolfl S, Hellwig FH (2004) The chloroplast genome of *Nymphaea alba*: whole-genome analyses and the problem of identifying the most basal angiosperm. Molec Biol Evol 21:1445–1454. doi:10.1093/molbev/msh147

Goremykin VV, Nikiforova SV, Biggs PJ, Zhong B, Delange P, Martin W, Woetzel S, Atherton RA, McLenachan PA, Lockhart PJ (2013) The evolutionary root of flowering plants. Syst Biol 62:50–61. doi:10.1093/sysbio/sys070

Goremykin VV, Nikiforova SV, Cavalieri D, Pindo M, Lockhart P (2015) The root of flowering plants and total evidence. Syst Biol 64:879–891. doi:10.1093/sysbio/syv028

Guisinger MM, Kuehl JV, Boore JL, Jansen RK (2011) Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. Molec Biol Evol 28:583–600. doi:10.1093/molbev/msq229

Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075. doi:10.1093/bioinformatics/btt086

Haberle RC, Fourcade HM, Boore JL, Jansen RK (2008) Extensive rearrangements in the chloroplast genome of *Trachelium caeruleum* are associated with repeats and tRNA genes. J Molec Evol 66:350–361. doi:10.1007/s00239-008-9086-4

Iles WJD, Lee C, Sokoloff DD, Remizowa MV, Yadav SR, Barrett MD, Barrett RL, Macfarlane TD, Rudall PJ, Graham SW (2014) Reconstructing the age and historical biogeography of the ancient flowering-plant family Hydatellaceae (Nymphaeales). BMC Evol Biol 14:102. doi:10.1186/1471-2148-14-102

Jansen RK, Ruhlman TA (2012) Plastid genomes of seed plants. In: Bock R, Knoop V (eds) Genomics of chloroplasts and mitochondria, advances in photosynthesis and respiration. Springer, Berlin, pp 103–126

Jansen RK, Cai Z, Raubeson LA, Daniell H, DePamphilis CW, Leebens-Mack J, Müller KF, Guisinger-Bellian M, Haberle RC, Hansen AK, Chumley TW, Lee S-B, Peery R, McNeal JR, Kuehl JV, Boore JL (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci 104:19369–19374. doi:10.1073/pnas.0709121104

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A (2012) Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–1649. doi:10.1093/bioinformatics/bts199

Kuroda H, Suzuki H, Kusumegi T, Hirose T, 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. Pl Cell Physiol 48:1374–1378. doi:10.1093/pcp/pcm097

Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B (2016) PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Molec Biol Evol 34:772. doi:10.1093/molbev/msw260

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25. doi:10.1186/gb-2009-10-3-r25

Leebens-Mack J, Raubeson LA, Liying C, Kuehl JV, Fourcade MH, Chhumley TW, Boore JL, Jansen RK, dePamphilis CW (2005) Identifying the basal angiosperm node in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone. Molec Biol Evol 22:1948–1963. doi:10.1093/molbev/msi191

Lewis PO (2001) A likelihood approach to estimating phylogeny from discrete morphological character data. Syst Biol 50:913–925. doi:10.1080/106351501753462876

Liu C, Shi L, Zhu Y, Chen H, Zhang J, Lin X, Guan X (2012) CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. BMC Genom 13:715. doi:10.1186/1471-2164-13-715

Logacheva MD, Schelkunov MI, Penin AA (2011) Sequencing and analysis of plastid genome in mycoheterotropic orchid *Neottia nidus-avis*. Genome Bio Evol 3:1296–1303. doi:10.1093/gbe/evr102

Löhne C, Borsch T (2005) Molecular evolution and phylogenetic utility of the petD group II intron: a case study in basal angiosperms. Molec Biol Evol 22:317–332. doi:10.1093/molbev/msi019

Löhne C, Borsch T, Wiersema JH (2007) Phylogenetic analysis of Nymphaeales using fast-evolving and noncoding chloroplast markers. Bot J Linn Soc 154:141–163. doi:10.1111/j.1095-8339.2007.00659.x

Löhne C, Yoo M-J, Borsch T, Wiersema J, Wilde V, Bell CD, Barthlott W, Soltis DE, Soltis PS (2008) Biogeography of Nymphaeales: extant patterns and historical events. Taxon 57:1123–1146

Löhne C, Wiersema JH, Borsch T (2009) The unusual *Ondinea*, actually just another Australian water-lily of *Nymphaea* subg. *Anecphya* (Nymphaeaceae). Willdenowia 39:55–58. doi:10.3372/wi.39.39104

Lohse M, Drechsel O, Kahlau S, Bock R (2013) OrganellarGenomeDRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression datasets. Nucl Acids Res 41:W575–W581. doi:10.1093/nar/gkt289

McNeal JR, Kuehl JV, Boore JL, de Pamphilis CW (2007) Complete plastid genome sequences suggest strong selection for retention of photosynthetic genes in the parasitic plant genus *Cuscuta*. BMC Pl Biol 7:57. doi:10.1186/1471-2229-7-57

Michaud M, Cognat V, Duchene A-M, Marechal-Drouard L (2011) A global picture of tRNA genes in plant genomes. Pl J 66:80–93. doi:10.1111/j.1365-313X.2011.04490.x

Moore MJ, Bell CD, Soltis PS, Soltis DE (2007) Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proc Natl Acad Sci USA 104:19363–19368. doi:10.1073/pnas.0708072104

Moore MJ, Hassan N, Gitzendanner MA, Bruenn RA, Croley M, Vandeventer A, Horn JW, Dhingra A, Brockington SF, Latvis M, Ramdial J, Alexandre R, Piedrahita A, Xi Z, Davis CC, Soltis PS, Soltis DE (2011) Phylogenetic analysis of the plastid inverted repeat for 244 species: insights into deeper-level angiosperm relationships from a long, slowly evolving sequence region. Int J Pl Sci 172:541–558. doi:10.1086/658923

Müller KF (2005) SeqState—primer design and sequence statistics for phylogenetic DNA datasets. Appl Bioinformatics 4:65–69. doi:10.2165/00822942-200504010-00008

Müller KF, Borsch T, Hilu KW (2006) Phylogenetic utility of rapidly evolving DNA at high taxonomical levels: contrasting *matK*, *trnT-F* and *rbcL* in basal angiosperms. Molec Phylogen Evol 41:99–117. doi:10.1016/j.ympev.2006.06.017

Müller J, Müller KF, Neinhuis C, Quandt D (2007) PhyDE—phylogenetic data editor. Available at: http://www.phyde.de

Ohtani K, Yamamoto H, Akimitsu K (2002) Sensitivity to *Alternaria alternata* toxin in citrus because of altered mitochondrial RNA processing. Proc Natl Acad Sci USA 99:2439–2444. doi:10.1073/pnas.042448499

Palmer JD (1983) Chloroplast DNA exists in two orientations. Nature 301:92–93. doi:10.1038/301092a0

Rambaut A, Suchard MA, Drummond AJ (2014) Tracer v1.6. Available at: http://beast.bio.ed.ac.uk/Tracer

Raubeson LA, Peery R, Chumley TW, Dziubek C, Fourcade HM, Boore JL, Jansen RK (2007) Comparative chloroplast genomics: analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. BMC Genom 8:174. doi:10.1186/1471-2164-8-174

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574. doi:10.1093/bioinformatics/btg180

Rudall PJ, Sokoloff DD, Remizowa MV, Conran JG, Davis JI, Macfarlane TD, Stevenson DW (2007) Morphology of Hydatellaceae, an anomalous aquatic family recently recognized as an early-divergent angiosperm lineage. Amer J Bot 94:1073–1092. doi:10.3732/ajb.94.7.1073

Ruhfel BR, Gitzendanner MA, Soltis PS, Soltis DE, Burleigh JG (2014) From algae to angiosperms—inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. BMC Evol Biol 14:23. doi:10.1186/1471-2148-14-23

Saarela JM, Rai HS, Doyle JA, Endress PK, Methews S, Marchant AD, Briggs BG, Graham SW (2007) Hydatellaceae identified as a new branch near the base of the angiosperm phylogenetic tree. Nature 446:312–315. doi:10.1038/nature05612

Schmitz-Linneweber C, Maier RM, Alcaraz J-P, Cottet A, Herrmann RG, Mache R (2001) The plastid chromosome of spinach (*Spinacia oleracea*): complete nucleotide sequence and gene organization. Pl Molec Biol 45:307–315. doi:10.1023/A:1006478403810

Schneider EL, Williamson PS (1993) Nymphaeaceae. In: Kubitzki K, Rohwer JG, Bittrich V (eds) Flowering plants—Dicotyledons: Magnoliid, Hamamelid and Caryophyllid Families. Springer, Berlin, pp 486–493

Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-Shinozaki K, Ohto C, Torazawa K, Meng BY, Sugital M, Deno H, Kamogashira T, Yamada K, Kusuda J, Takaiwa F, Kato A, Tohdoh N, Shimada H, Sugiura M (1986) The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J 5:2043–2049

Silva SR, Diaz YCA, Penha HA, Pinheiro DG, Fernandes CC, Miranda VFO, Michael TP, Varani AM (2016) The chloroplast genome of *Utricularia reniformis* sheds light on the evolution of the *ndh* gene complex of terrestrial carnivorous plants from the Lentibulariaceae family. PLoS ONE 11:1–29. doi:10.1371/journal.pone.0165176

Simmons MP (2016) Mutually exclusive phylogenomic inferences at the root of the angiosperms: *Amborella* is supported as sister and observed variability is biased. Cladistics (First Online). doi:10.1111/cla.12177

Simmons MP, Ochoterena H (2000) Gaps as characters in sequence-based phylogenetic analyses. Syst Biol 49:369–381

Soininen EM, Valentini A, Coissac E, Miquel C, Gielly L, Brochmann C, Brysting AK, Sonstebo JH, Ims RA, Yoccoz NG, Taberlet P (2009) Analysing diet of small herbivores: the efficiency of DNA barcoding coupled with high-throughput pyrosequencing for deciphering the composition of complex plant mixtures. Frontiers Zool 6:16. doi:10.1186/1742-9994-6-16

Soltis PS, Soltis DE, Chase MW (1999) Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature 402:402–404. doi:10.1038/46528

Stamatakis A (2014) RAxML Version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. doi:10.1093/bioinformatics/btu033

Stamatakis A, Hoover P, Rougemont J (2008) A rapid bootstrap algorithm for the RAxML Web servers. Syst Biol 57:758–771. doi:10.1080/10635150802429642

Stevens PF (2017) Angiosperm phylogeny website. Version 13 [last update 05/02/2017]. Available at: http://www.mobot.org/MOBOT/research/APweb/

Sun L, Fang L, Zhang Z, Chang X, Penny D (2016) Chloroplast phylogenomic inference of green algae relationships. Sci Rep 6:20528. doi:10.1038/srep20528

Walker JF, Jansen RK, Zanis MJ, Emery NC (2015) Sources of inversion variation in the small single copy (SSC) region of chloroplast genomes. Amer J Bot 102:1751–1752. doi:10.3732/ajb.1500299

Weng M-L, Blazier JC, Govindu M, Jansen RK (2014) Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats, and nucleotide substitution rates. Molec Biol Evol 31:645–659. doi:10.1093/molbev/mst257

Wicke S, Schneeweiss G, dePamphilis C, Müller K, Quandt D (2011) The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Pl Molec Biol 76:273–297. doi:10.1007/s11103-011-9762-4

Wicke S, Muller KF, de Pamphilis CW, Quandt D, Wickett NJ (2013) Mechanisms of functional and physical genome reduction in photosynthetic and nonphotosynthetic parasitic plants of the broomrape family. Pl Cell 25:3711–3725. doi:10.1105/tpc.113.113373

Wicke S, Schäferhoff B, Depamphilis CW, Müller KF (2014) Disproportional plastome-wide increase of substitution rates and relaxed purifying selection in genes of carnivorous Lentibulariaceae. Molec Biol Evol 31:529–545. doi:10.1093/molbev/mst261

Woloszynska M, Bocer T, Mackiewicz P, Janska H (2004) A fragment of chloroplast DNA was transferred horizontally, probably from non-eudicots, to mitochondrial genome of

Phaseolus. Pl Molec Biol 56:811–820. doi:10.1007/s11103-004-5183-y

Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20:3252–3255. doi:10.1093/bioinformatics/bth352

Yang J-B, Li D-Z, Li H-T (2014) Highly effective sequencing whole chloroplast genomes of angiosperms by nine novel universal primer pairs. Molec Ecol Resources 14:1024–1031. doi:10.1111/1755-0998.12251

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829. doi:10.1101/gr.074492.107

Zhang J, Ruhlman TJ, Sabir JSM, Blazier JC, Weng M-L, Park S, Jansen RK (2016) Coevolution between nuclear-encoded DNA replication, recombination, and repair genes and plastid genome complexity. Genome Biol Evol 8:622–634. doi:10.1093/gbe/evw033