

Fort Hays State University

FHSU Scholars Repository

Biological Sciences Faculty Publications

Faculty Publications

5-18-2024

On the importance of sequence alignment inspections in plastid phylogenomics – an example from revisiting the relationships of the water-lilies

Jessica A. Roestel

John H. Wiersema

Robert K. Jansen

Thomas Borsch


Michael Gruenstaeudl

Follow this and additional works at: https://scholars.fhsu.edu/biology_facpubs



Part of the [Biology Commons](#), and the [Botany Commons](#)

On the importance of sequence alignment inspections in plastid phylogenomics – an example from revisiting the relationships of the water-lilies

Jessica A. Roestel^a, John H. Wiersema^b, Robert K. Jansen^c, Thomas Borsch^{a,d} and Michael Gruenstaeudl^{*a,c} 

^aInstitut für Biologie, Systematische Botanik und Pflanzengeographie, Freie Universität Berlin, Berlin 14195, Germany; ^bDepartment of Botany, National Museum of Natural History – Smithsonian Institution, Washington, DC 37012, USA; ^cDepartment of Integrative Biology, University of Texas at Austin, Austin, TX 78712, USA; ^dBotanischer Garten und Botanisches Museum Berlin, Freie Universität Berlin, 14195 Berlin, Germany; ^eDepartment of Biological Sciences, Fort Hays State University, Hays, KS 67601, USA

Received 9 January 2024; Revised 27 April 2024; Accepted 29 April 2024

Abstract

The water-lily clade represents the second earliest-diverging branch of angiosperms. Most of its species belong to Nymphaeaceae, of which the “core Nymphaeaceae”—comprising the genera *Euryale*, *Nymphaea* and *Victoria*—is the most diverse clade. Despite previous molecular phylogenetic studies on the core Nymphaeaceae, various aspects of their evolutionary relationships have remained unresolved. The length-variable introns and intergenic spacers are known to contain most of the sequence variability within the water-lily plastomes. Despite the challenges with multiple sequence alignment, any new molecular phylogenetic investigation on the core Nymphaeaceae should focus on these noncoding plastome regions. For example, a new plastid phylogenomic study on the core Nymphaeaceae should generate DNA sequence alignments of all plastid introns and intergenic spacers based on the principle of conserved sequence motifs. In this investigation, we revisit the phylogenetic history of the core Nymphaeaceae by employing such an approach. Specifically, we use a plastid phylogenomic analysis strategy in which all coding and noncoding partitions are separated and then undergo software-driven DNA sequence alignment, followed by a motif-based alignment inspection and adjustment. This approach allows us to increase the reliability of the character base compared to the default practice of aligning complete plastomes through software algorithms alone. Our approach produces significantly different phylogenetic tree reconstructions for several of the plastome regions under study. The results of these reconstructions underscore that *Nymphaea* is paraphyletic in its current circumscription, that each of the five subgenera of *Nymphaea* is monophyletic, and that the subgenus *Nymphaea* is sister to all other subgenera of *Nymphaea*. Our results also clarify many evolutionary relationships within the *Nymphaea* subgenera *Brachyceras*, *Hydrocallis* and *Nymphaea*. In closing, we discuss whether the phylogenetic reconstructions obtained through our motif-based alignment adjustments are in line with morphological evidence on water-lily evolution.

© 2024 The Authors. *Cladistics* published by John Wiley & Sons Ltd on behalf of Willi Hennig Society.

Introduction

Systematics and diversity of Nymphaeaceae

Nymphaeaceae are a near-cosmopolitan family of flowering plants that comprise approximately 75 species of aquatic herbs in five distinct genera

(i.e. *Barclaya*, *Euryale*, *Nuphar*, *Nymphaea* and *Victoria*; Borsch et al., 2008; Löhne et al., 2009). *Nymphaea*, commonly referred to as “water-lilies,” represents the largest and most widely distributed genus of the family. The family is sister to the largely Neotropical Cabombaceae, which include only six or seven species in two genera (*Brasenia* and *Cabomba*; Löhne et al., 2007; Barbosa et al., 2018). Ecologically, Nymphaeaceae are highly diverse and inhabit many freshwater habitats of the temperate, subtropical, and

*Corresponding author:

E-mail address: m_gruenstaeudl@fhsu.edu

tropical zones (Löhne et al., 2007). The primary centres of diversity of Nymphaeaceae correlate with the distribution areas of three of the five subgenera of *Nymphaea*: northern South America, the Caribbean, the Zambezi region of Africa, and northern Australia (Löhne et al., 2008b).

The largest clade of Nymphaeaceae comprises the three genera *Nymphaea*, *Euryale*, and *Victoria* and is often referred to as the “core Nymphaeaceae” (Borsch et al., 2007)—a terminology we follow here as well. It has been recovered in several molecular phylogenetic investigations with maximum node support (e.g. Borsch et al., 2007; Löhne et al., 2007). Morphologically, this clade is supported by a protruding floral axis that distinctly exceeds the surrounding carpel tissue (Moseley, 1961; Borsch et al., 2008), a tetramerous outer perianth (Schneider et al., 2003), the presence of an aril and stipules (Borsch et al., 2008), and floral vasculature characters (Moseley et al., 1993; Les et al., 1999). Species diversity within the core Nymphaeaceae is highly unequal: *Nymphaea* comprises approximately 55–60 extant species (Borsch et al., 2011), whereas the genera *Euryale* and *Victoria* include only one and three species, respectively (Smith et al., 2020). Originally, the core Nymphaeaceae also included the monotypic Australian genus *Ondinea*, but its species (*O. purpurea*) has been identified as a member of *Nymphaea* subgen. *Anecphyta* and it was transferred into the genus as *N. ondinea* (Löhne et al., 2009). *Nymphaea* has been subdivided into five subgenera that reflect their primary distribution areas: the Papuan-Australian *Anecphyta* (c. 16 species), the pantropical *Brachyceras* (c. 16 species), the Neotropical *Hydrocallis* (c. 15 species), the Palaeotropical *Lotos* (c. three species), and the north-temperate *Nymphaea* (c. eight species; Borsch et al., 2007; Löhne et al., 2007).

Three reasons render the exploration of the evolutionary history of Nymphaeaceae important. First, the family represents one of the earliest-diverging lineages within the flowering plants, making it relevant for understanding the history of early angiosperms (Zhang et al., 2020). Specifically, the morphological, cytological and spatiotemporal evolution of this lineage is relevant for evaluating hypotheses on the ancestor of all angiosperms (Borsch et al., 2008). Second, the evolution of Nymphaeaceae is of high interest for ecological and morphological aspects. For example, their evolutionary history indicates that Nymphaeaceae have developed adaptations to seasonal habitats and different pollination strategies (Löhne et al., 2008b). Third, our knowledge of the phylogenetic relationships of Nymphaeaceae and relatives has remained fragmentary. Specifically, the radiation of the core Nymphaeaceae is only partially understood. For example, the exact position of *Victoria* within *Nymphaea* and the relationships within the clade formed by *Nymphaea* subgenera *Anecphyta*

and *Brachyceras* have yet to be clarified (Löhne et al., 2008b).

Previous molecular phylogenetic investigations on Nymphaeaceae

Several molecular phylogenetic investigations have explored the evolutionary history within and among genera of Nymphaeaceae. Most of these studies employed traditional Sanger sequencing to generate nucleotide sequence data for one or a few genomic loci. Les et al. (1999), for example, evaluated the phylogenetic relationships between different genera of water-lilies using both nuclear and plastid genome loci. A dataset with a denser representation of *Nymphaea* was compiled and evaluated by Borsch et al. (2007), who compared interspecific relationships in *Nymphaea* with the plastid *trnT-trnF* region and investigated the molecular evolution of its two spacers and the group I intron in *trnL*. Their results indicated weak to medium support for the monophyly of *Nymphaea* and the presence of three major lineages within the genus: a strongly supported clade comprising the subgenera *Hydrocallis* and *Lotos*, which share various morphological synapomorphies (e.g. night-blooming flowers predominantly pollinated by beetles); a clade comprising subgenera *Anecphyta* and *Brachyceras*; and a clade comprising subgenus *Nymphaea*, which was found to be sister to a clade comprising all other species of the genus. Their study was also the first to indicate that the genus *Ondinea* was nested within *Nymphaea* and a close relative of subgenus *Anecphyta*. Likewise, Löhne et al. (2007) analysed different plastome regions among 12 members of Nymphaeaceae, including all subgenera of *Nymphaea*. Their results indicated that *Nymphaea* may be paraphyletic unless the genera *Ondinea*, *Victoria* and *Euryale* were included, that *Ondinea* formed a clade with members of subgenus *Anecphyta*, and that *Victoria* and *Euryale* were closely related to subgenera *Hydrocallis* and *Lotos*.

Borsch et al. (2008) extended the taxon sampling of previous studies to members of the Cabombaceae and analysed DNA sequences of the nuclear ribosomal internal transcribed spacer (nrITS) region and the mitochondrial gene *matR* across members of each genus and contrasted their results with morphological data. They found that Nymphaeaceae were monophyletic, and that the genera *Nuphar* and *Barclaya* were successive sister taxa to the core Nymphaeaceae. Löhne et al. (2008a) extended the taxon sampling of previous investigations to generate the first near-comprehensive analysis of phylogenetic relationships among the Australian water-lilies (*Nymphaea* subgenus *Anecphyta*). They found that subgenus *Anecphyta* is split into two clades that exhibit different seed size and that the monotypic genus *Ondinea* has a close

relationship to members of the small-seeded clade of this subgenus. This phylogenetic placement of *Ondinea* was taxonomically formalized by Löhne et al. (2009) when they subsumed the genus under *Nymphaea* as *N. ondinea*. Löhne et al. (2008a) detected phylogenetic incongruence between plastid and nuclear genome sequence data, which suggested hybridization or introgression among the Australian species of *Nymphaea*. Borsch et al. (2011) analysed nrITS and plastid *trnT-trnF* DNA sequences to clarify the relationships between water-lilies of subgenus *Brachyceras* and the Australian representatives of *Nymphaea*. Their results indicated the monophyly of subgenus *Brachyceras*, which includes the presence of a clade of American species that are phylogenetically nested among some of their African consubgenera. Additional attention to the North American species of *Nymphaea* was given by Borsch et al. (2014), who analysed the plastid *trnT-trnF* region among 43 samples of subgenus *Nymphaea* and found that *N. leibergii* and *N. tetragona* were likely to be sister species.

Previous plastid phylogenomic studies of Nymphaeaceae

The phylogenetic position and the monophyly of Nymphaeaceae as well as the size and structure of their plastid genomes have been evaluated in several previous investigations. The first study to phylogenetically analyse and compare water-lily plastomes was conducted by Gruenstaeudl et al. (2017), who identified conservation in size and gene content, strong support for a sister relationship between Nymphaeaceae and Cabombaceae, and strong support for the monophyly of the Cabombaceae. However, only weak, if any, support was obtained for the monophyly of Nymphaeaceae because *Nuphar* was repeatedly recovered as sister to either Cabombaceae alone or a clade of the Cabombaceae and the remainder of Nymphaeaceae. He et al. (2018) retrieved and supplemented the taxon set of Gruenstaeudl et al. (2017) in a second investigation on the monophyly of Nymphaeaceae. While their results indicated that Nymphaeaceae were monophyletic, a subsequent gene-wise re-evaluation of their analyses found the monophyly of Nymphaeaceae remained unresolved (Gruenstaeudl, 2019). Nonetheless, each of these studies supported the monophyly of the core Nymphaeaceae, but many of the intergeneric and especially interspecific relationships have remained uncertain.

Importance of motif-based sequence alignment in plastid phylogenomics

Although the positional homology among the nucleotides of a DNA sequence alignment represents the data foundation for molecular phylogenetic inferences

(Ogden and Rosenberg, 2006), not all phylogenetic investigations pay sufficient attention to the accuracy of these alignments. In phylogenetics, a DNA sequence alignment is a set of assumptions on the homology between the individual nucleotides of input DNA sequences and typically inferred through the process of multiple sequence alignment (MSA; Phillips et al., 2000). Homology assumptions across nucleotide sequences are mostly grounded in molecular genetic theories that were established based on the empirical observation of DNA sequence mutations over time (Morrison, 2006). As such, DNA sequence alignments are inferences that can be more or less accurate, depending on how well the underlying mutational dynamics are represented through the implied positional homologies (Kelchner, 2000; Borsch and Quandt, 2009). The use of sequence motifs that are conserved—and that can, thus, be tracked—across lineages has been found particularly beneficial when trying to identify homologous regions across DNA sequences and, by extension, to conduct MSA (Hickson et al., 2000; Morrison, 2008). Conserved sequence motifs are nucleotide regions that share one or more structural or functional constraints and are maintained across different organismic lineages. Prominent examples of conserved sequence motifs are gene promoters, the binding sites of transcription factors, and hairpin loops representing secondary DNA structure (Hickson et al., 1996). Owing to their structural or functional constraints, conserved sequence motifs do not typically experience random nucleotide substitutions but instead exhibit mutations that are in line with the constraints of the motifs (e.g. Liang et al., 2018; Pereira et al., 2022). Moreover, plastome DNA exhibits characteristic patterns of microstructural mutations that typically affect multiple nucleotides at once, such as the expansion or contraction of simple sequence repeats (SSRs) or the inversion of entire sequence elements (Kelchner, 2000; Morrison, 2008; Borsch and Quandt, 2009). Thus, sequence motifs do not conform to the implicit assumption of most alignment algorithms, which weigh each sequence position equally and assume gaps of individual nucleotides as fifth character states. Genomic sequences that contain numerous conserved sequence motifs represent a mosaic of evolutionary pressures that can be challenging to model algorithmically (Chatzou et al., 2016). Hence, the process of MSA of genome regions with an abundance of conserved sequence motifs requires a different alignment strategy than the MSA of largely unconstrained genome regions (Dijkstra et al., 2018).

Over the past decades, various studies have highlighted the negative effect that imprecise sequence alignments—and, thus, incorrect assumptions of positional homology—can have on the accuracy of phylogenetic tree inference (e.g. Wong et al., 2008). Despite

these findings, many phylogenetic investigations, including most phylogenomic ones, assume that the positional homology established through automatic, software-driven sequence alignment is either correct by default or at least repeatable and, thus, suitable for phylogenetic inference (Morrison, 2006, 2009). While this simplification may be sufficient in cases where only the coding regions of genomes are phylogenetically analysed, the positional homologies of which are easier to establish (e.g. Leebens-Mack et al., 2005), it can be misleading when noncoding genome regions are compared (e.g. Escobari et al., 2021). The plastid genome of land plants, for example, comprises a large proportion of noncoding DNA regions, which constitutes 40–45% of the overall genome length and typically exhibits a higher average rate of nucleotide substitution than the coding regions (Shaw et al., 2007). In most plastid phylogenomic datasets, the noncoding regions, thus, contain the majority of all potentially informative characters (e.g. Korotkova et al., 2014). Next to nucleotide substitutions, the intron and spacer regions of plastid genomes, and even some of their genes (e.g. *matK*; Hilu et al., 2003), can accumulate small microstructural mutations, including SSRs, insertions and deletions, and short sequence inversions (Graham et al., 2000; Kelchner, 2000; Borsch and Quandt, 2009). Some AT-rich intron and spacer regions with SSRs or minisatellites may even be hypervariable, resulting in a virtual inability to establish positional homology for these regions (Borsch et al., 2003; Korotkova et al., 2014). Dedicated MSA strategies are, thus, critical for phylogenomic analyses of noncoding sequence data (e.g. Escobari et al., 2021).

The higher frequency of microstructural mutations in noncoding compared to coding plastome regions warrants extra assessments of the positional homology in plastid phylogenomic data (Kelchner, 2000; Löhne and Borsch, 2005). Such assessments typically comprise an inspection and, where necessary, a manual adjustment of the software-generated sequence alignments (Morrison, 2006; Escobari et al., 2021). Alignment adjustments that are guided by conserved sequence motifs have been found to be particularly beneficial in this process (Löhne and Borsch, 2005; Morrison, 2008), as they measurably reduced the levels of homoplasy in the aligned regions (Escobari et al., 2021). Before the widespread application of phylogenomic datasets (i.e. when only a few genomic regions were analysed in molecular phylogenetic studies), such alignment inspections were commonplace (Morrison et al., 2015). In contemporary molecular phylogenetics, however, such inspections have largely fallen out of favour, not least as a consequence of their perceived time intensity given the size of most phylogenomic datasets (e.g. Wu et al., 2012).

Most phylogenomic investigations instead apply automated remedies against uncertain sequence alignments, such as dynamic gap penalties during automated MSA (e.g. Sela et al., 2015) or the masking of alignment regions with suboptimal confidence scores (e.g. Wu et al., 2012). However, these strategies typically ignore the underlying motif-based sequence evolution (Morrison, 2009), and some have even been found to reduce the accuracy of the phylogeny inference (e.g. Tan et al., 2015). Especially in cases where most species under study have diverged recently and the clades are characterized by short internal nodes or shallow terminal subclades, which are the very evolutionary conditions when the comparison of complete plastid genomes is deemed the most useful (e.g. Escobari et al., 2021), the application of alignment inspection and adjustment is rarely employed.

Given the ample evidence on the importance of accurate MSA for accurate phylogenetic reconstruction (Ogden and Rosenberg, 2006), plastid phylogenomic investigations should not shy away from the inspection and, where necessary, motif-aware adjustment of software-generated sequence alignments, even if such inspections appear time-consuming owing to the large amounts of sequence data involved. In fact, in phylogenomic studies of taxon groups with low genetic distances, motif-based alignment adjustments may have the strongest impact on phylogenetic accuracy (Smith et al., 2020). Because the genetic distances among the genera of Nymphaeaceae are generally low (Gruenstaeudl et al., 2017), any plastid phylogenomic investigation of this plant family should place an emphasis on a strategy of motif-based alignment inspection and adjustment with the aim of improving the reliability of the reconstruction results.

Aims

Here, we conduct a plastid phylogenomic investigation of Nymphaeaceae with a species sampling of *Nymphaea* that is taxonomically greater than previous plastid phylogenetic studies of this group. At the heart of our study is a sequence alignment approach that includes the visual inspection and the motif-based adjustment of software-generated DNA sequence alignments. Specifically, we sequence and assemble the complete plastomes of 21 species of *Nymphaea* and one species of *Barclaya*, combine them with nine previously published plastid genomes, partition this plastome set by gene, intron and intergenic spacer, align all partitions bioinformatically, and then manually inspect and, where necessary, adjust the alignments using a motif-based evaluation approach, followed by phylogenetic tree inference. We then assess if our alignment adjustments improve the positional homology among the aligned sequences, affect the inference

of the best-fitting nucleotide substitution models, and impact upon the phylogenetic tree reconstructions. Additional analyses regarding the coding of small nucleotide insertions/deletions and the presence of small sequence inversions among the aligned sequences are also conducted. Taken together, we reconstruct the phylogenetic history of Nymphaeaceae from complete plastid genome sequences despite the challenges posed by the idiosyncratic evolution of the coding and non-coding plastome regions, as well as the low genetic distances between the members of this clade, and assess if the motif-based post-processing of software-generated sequence alignments measurably improves the reliability of these reconstructions.

Materials and methods

Taxon sampling

Plant samples of 21 taxa of Nymphaeaceae, representing 19 different species, were collected and sequenced for this investigation. Specifically, we collected 18 samples of *Nymphaea*, representing each of the five subgenera (i.e. subgenera *Anecphyia*, *Brachyceras*, *Hydrocallis*, *Lotos* and *Nymphaea*), two species of *Nuphar* and one species of *Barclaya*. Among the samples of *Nymphaea* are two accessions of *N. glandulifera* and two accessions of *N. lotus*, both of which represent different geographical origins of these species. All samples were collected either as young leaves from live plant specimens cultivated at the Botanical Garden and Botanical Museum Berlin (BGBM) or obtained as deep-frozen leaf material stored at -80°C from previous fieldwork (see Table 1 for voucher information). Next to these new plastid genomes, we also included several previously published plastid genomes in our analyses but excluded the genomes of plants whose taxonomic identity was uncertain. Specifically, we only included plastid genomes that were generated from documented plant material of wild populations and had specimen vouchers available in public herbaria. By comparison, we abstained from including plastid genomes that lacked references to herbarium vouchers or other identifying physical material in their source publications. Given that the species limits among water-lilies are not clear in all cases and that many plants in botanical gardens represent ornamental hybrids, we also tried to avoid the integration of redundant plastid genomes or those with an uncertain taxonomic identification into our final dataset. This conservative selection process resulted in the integration of nine previously published plastome sequences from four different genera into our analyses: we included the plastid genomes of three species of *Nymphaea* (*N. alba*, *N. ampla* and *N. jamesoniana*; all Gruenstaeudl et al., 2017), three species of *Victoria* (*V. cruziana*—Gruenstaeudl et al., 2017 and Smith et al., 2022; *V. amazonica* and *V. boliviana*—both Smith et al., 2022), one species of *Barclaya* (*B. longifolia*; Gruenstaeudl et al., 2017) and the outgroup species *Cabomba caroliniana* (NC_031505; Gruenstaeudl et al., 2017) into a combined dataset. The final dataset, thus, comprised 30 complete plastomes that represent a total of 22 different species of recognized genera of the core Nymphaeaceae, four species of the family outside the core Nymphaeaceae, and a representative of the sister family to Nymphaeaceae (i.e. *Cabomba* of the Cabombaceae). A complete list of the newly sequenced as well as the previously published plastomes, species names, the taxonomic authorities of these species, herbarium voucher and DNA isolate identifiers, taxonomic affiliation to subgenus of *Nymphaea*, and GenBank and NCBI SRA accession numbers is given in Table 1.

DNA extraction and Illumina sequencing

For each collected plant sample, total genomic DNA was extracted and sequenced using a genome-skimming approach. To extract DNA, young leaves were surface-cleaned with deionized water and 70% ethanol, and desiccated on silica gel for 24 h. Total genomic DNA was isolated from 20 mg of dried leaf material using a modified version of the CTAB isolation method (Borsch et al., 2003). Extracted DNA was purified using the DNeasy PowerClean Pro Cleanup Kit (Qiagen, Hilden, Germany) and then sheared via ultrasonication to an average fragment size of ~ 300 bp. Concentration and fragment distribution of the sheared DNA were measured on a Fragment Analyser System 5200 (Agilent Technologies, Santa Clara, CA, USA). Upon confirming optimal fragment size, a barcoded DNA library was constructed for each sample using the TruSeq DNA samples preparation kit (Illumina, San Diego, CA, USA). The libraries were pooled equimolarly and sequenced as paired end reads on an Illumina HiSeq X platform (Illumina, San Diego, CA, USA) by Macrogen (Seoul, Republic of Korea).

Plastome assembly and annotation

After DNA sequencing, adapter sequences were trimmed from the reads, and reads with low quality scores were removed from the read set using Trimmomatic v.0.36 (Bolger et al., 2014). The resulting reads were mapped to the five previously published plastomes of Nymphaeaceae generated by Gruenstaeudl et al. (2017) using script #5 of the pipeline of Gruenstaeudl et al. (2018) to extract all plastome reads and, subsequently, conduct genome assembly on plastome reads only. After read extraction, complete plastomes were assembled *de novo* with the software GetOrganelle v.1.6.4 (Jin et al., 2020). Recently, Giorgashvili et al. (2022) tested the performance of different software tools for plastome assembly under different coverage levels, and concluded that GetOrganelle generated the most consistent and reliable assemblies under a sequencing coverage of $\times 100$ to $\times 500$. If this initial assembly with GetOrganelle did not result in a complete plastome, we conducted an additional *de novo* assembly using NOVOPlasty v.3.8.3 (Dierckxens et al., 2017) and then compared and resolved the contigs of both assembly tools visually in Geneious v.11.1.4 (Kearse et al., 2012). We manually standardized the orientation of the single-copy regions across all new plastid genomes to assist the automated extraction of coding regions across the plastid genomes. Upon assembly, a visual examination of the resulting sequences was conducted to identify segments of suboptimal assembly; any such segments were masked as question mark characters during alignment adjustment to avoid spurious phylogenetic signal. For example, previously generated Sanger sequences of the AT-rich stem loop of the *trnL* group I intron (Borsch et al., 2007) were used as template sequences to identify segments of suboptimal assembly. A summary of these alignment adjustments is listed in Table S1. To annotate the new plastomes, all protein-coding gene, tRNA and rRNA annotations of five previously published plastid genomes of Nymphaeaceae were transferred to the new genomes using Geneious under a sequence similarity threshold of 95%. Upon transfer, the annotations were visually inspected and, where necessary, corrected regarding the presence of start and stop codons and the absence of internal stop codons. Final plastome sequences were deposited in GenBank and their accession numbers listed in Table 1. The circular representation of the plastome of *N. immutabilis* was generated with OGDRAW v.1.3.1 (Greiner et al., 2019).

Combining plastomes into genome set

Before combining previously published and newly sequenced plastomes into a single genome set, we compared and corrected the

Table 1

Species name, taxonomic authority, GenBank and NCBI SRA accession number, DNA isolate identifier, and herbarium voucher for each plastid genome under study

	Taxonomic authority	GenBank accession	NCBI SRA accession	DNA isolate identifier	Herbarium voucher
<i>Barclaya longifolia</i>	Wall.	KY284156*	n.a.	n.a.	Gartenherbar Cubr 49678 (B)
<i>Barclaya rotundifolia</i>	Hotta	MW057721	PRJNA665362	DB 40839	Gartenherbar Cubr 50705 (B) [Indonesia: Wongso & Ipor s.n.]
<i>Cabomba caroliniana</i>	A. Grey	NC_031505*	n.a.	n.a.	J.C. Ludwig s.n. (VPI)
<i>Nuphar lutea</i>	(L.) Sm.	MH161175	PRJNA648088	DB 40167	Gartenherbar Cubr 50790 (B)
<i>Nuphar pumila</i>	(Timm) DC.	MH161176	PRJNA643573	DB 40168	Gartenherbar 50791 (B)
Subgenus <i>Anechpaya</i> (Casp.) Conard					
<i>Nymphaea immutabilis</i>	S.W.L. Jacobs	MW057732	PRJNA662618	DB 40760	Gartenherbar Cubr 51902 (B) [Australia: Jacobs s.n. BONN 19890]
Subgenus <i>Brachyceras</i> (Casp.) Conard					
<i>Nymphaea ampla</i>	(Salisb.) DC.	KU189255*	n.a.	n.a.	Gartenherbar Cubr 48929 (B)
<i>Nymphaea cf. capensis</i>	Thunb.	MW057740	PRJNA663178	DB 40834	Gartenherbar Cubr 51903 (B) [Botswana]
<i>Nymphaea dimorpha</i>	I.M. Turner	MW057738	PRJNA662385	DB 40735	Gartenherbar Cubr 51202 (B) [Madagascar: Anonymous s.n.]
<i>Nymphaea gracilis</i>	Zucc.	MW057734	PRJNA662360	DB 40598	A. Novelo R., J.H. Wiersema, C.B. Hellquist & C.B. Horn 1314 (MEXU)
<i>Nymphaea heudelotii</i>	Planch.	MW057733	PRJNA664003	DB 40838	Gartenherbar Cubr 42572 (B) [Rwanda: Fischer 3036]
<i>Nymphaea thermarum</i>	Eb. Fisch.	MW057722	PRJNA661718	DB 40234	Gartenherbar Cubr 51901 (B) [Rwanda: E. Fischer s.n.]
<i>Nymphaea × daubenyana</i>	W.T. Baxter ex Daubeny	MW057739	PRJNA666060	DB 40165	Gartenherbar Cubr 36074 (B) from cultivated source
Subgenus <i>Hydrocallis</i> (Planch.) Conard					
<i>Nymphaea amazonum</i>	Mart. & Zucc.	MW057741	PRJNA662994	DB 40833	Gartenherbar Cubr 51223 (B) [French-Guayana: N. Köster 2896]
<i>Nymphaea conardii</i>	Wiersema	MW057737	PRJNA659339	DB 40191	A. Novelo R., J.H. Wiersema, C.B. Hellquist & C.B. Horn 1306 (MEXU)
<i>Nymphaea glandulifera</i>	Rodschied	MW057735	PRJNA666444	DB 40758	Gartenherbar Cubr 51905 (B)
<i>Nymphaea glandulifera</i>	Rodschied	MW057736	PRJNA661714	DB 40559	C.N. Horn & J.H. Wiersema 4523 (US, BRG, NBYC)
<i>Nymphaea jamesoniana</i>	Planch.	KT749898*	n.a.	n.a.	T. Borsch & B. Summers 3218 (B)
<i>Nymphaea lasiophylla</i>	Mart. & Zucc.	MW057731	PRJNA662580	DB 40759	Gartenherbar Cubr 51265 (B) [Venezuela]
<i>Nymphaea rudgeana</i>	G. Mey.	MW057725	PRJNA662488	DB 40757	Gartenherbar Cubr 51214 (B) [Brazil: Anonymous s.n.]
<i>Nymphaea tenerinervia</i>	Casp.	MW057723	PRJNA661560	DB 40309	C.N. Horn & J.H. Wiersema 11086 (US, BRG, NBYC)
Subgenus <i>Lotos</i> (DC.) Conard					
<i>Nymphaea lotus</i>	L.	MW057729	PRJNA666778	DB 40836	Gartenherbar Cubr 23866 (B) [Togo: Anonymous s.n.]
<i>Nymphaea lotus</i>	L.	MW057730	PRJNA663224	DB 40835	Gartenherbar Cubr 38600 (B) [Ungarn: Anonymous s.n.]
Subgenus <i>Nymphaea</i>					
<i>Nymphaea alba</i>	L.	KU234277*	n.a.	n.a.	T. Borsch (B) [Italy, Lake Iseo]
<i>Nymphaea mexicana</i>	Zucc.	MW057727	PRJNA660901	DB 40239	T. Borsch & B. Summers 3227 (B, VPI) same population as NY069
<i>Nymphaea odorata</i>	Aiton	MW057726	PRJNA659634	DB 40211	T. Borsch & V. Wilde 3099 (B, VPI)
<i>Victoria amazonica</i>	(Poepp.) J.C. Sowerby	Unknown**	n.a.	n.a.	Te et al. 137 (Adelaide)
<i>Victoria boliviana</i>	Magdalena & L.T. Sm.	Unknown**	n.a.	n.a.	Magdalena et al. 1 (USZ)
<i>Victoria cruziana</i>	A.D. Orb.	KY001813*	n.a.	n.a.	C. Loehne 55 (BONN)
<i>Victoria cruziana</i>	A.D. Orb.	Unknown**	n.a.	n.a.	Sparre & Vervoorst 2363 (P)

Samples of *Nymphaea* are sorted by subgenus. Taxonomic authorities are provided in their abbreviated form following the abbreviation style of the International Plant Names Index. Sequences were originally published in: *, Gruenstaeudl et al. (2017); **, Smith et al. (2022). Abbreviations used: n.a. = not applicable.

sequence annotations of all previously published plastomes included in this investigation. No annotation differences were detected among any of the plastomes of Gruenstaeudl et al. (2017), but several were detected among those of Smith et al. (2022). These annotation conflicts were resolved based on the inferences of the annotation service GeSeq v.2.03 (Tillich et al., 2017). In preparation of sequence extraction and alignment, we bioinformatically removed the second of the two IRs from each genome to avoid redundancy among the extracted sequences.

Extraction and sequence alignment of coding regions

In order to generate sequence matrices of the coding plastome regions, the protein-coding sections of all plastid genomes were bioinformatically extracted, grouped by gene name, and aligned based on their amino acid sequence using a set of Python scripts provisionally termed “PlastomeBurstAndAlign” (<https://github.com/michaelgruenstaeudl/PlastomeBurstAndAlign>). Specifically, the 79 protein-coding genes found among all water-lily plastomes were extracted, translated from nucleotides to amino acids, aligned with MAFFT v.7.471 (Kato and Standley, 2013) under its default algorithm and parameter settings, and translated back to nucleotides using PlastomeBurstAndAlign. The default algorithm and parameter settings of MAFFT were found to be the best tradeoff between alignment accuracy and alignment speed in a preliminary analysis of this investigation (data not shown). We found that neither the sequence alignments generated by any of the iterative refinement algorithms of MAFFT nor the alignments generated by its default progressive algorithm accounted for complex microstructural mutations such as small sequence inversions, whereas differences in parameterization rendered the progressive algorithm often twice as fast as the iterative refinement algorithms (see also Long et al., 2016). Alignments of the coding regions were conducted in a gene-wise fashion owing to the advantage that any nucleotide insertion or deletion at the start or end of a coding region cannot cause an overlap of adjacent genes upon alignment. Likewise, alignments based on amino acids have the advantage that any insertion or deletion will automatically constitute a multiple of three and, thus, maintain the reading-frame (reviewed in Gruenstaeudl et al., 2018). Of the 79 extracted protein-coding regions, *rps12* was removed from further analysis because this gene is trans-spliced and contains discontinuous group II introns, which increases the error risk during bioinformatic processing. The alignments of the remaining 78 protein-coding genes were inspected for optimal positional homology and, where necessary, adjusted to improve alignment as discussed below.

Extraction and sequence alignment of noncoding regions

In order to generate sequence matrices of the noncoding plastome regions for phylogenetic inference, the intergenic spacers and introns of the plastid genomes were bioinformatically extracted, grouped by name, and aligned based on their nucleotide sequence using PlastomeBurstAndAlign. Specifically, a total of 110 intergenic spacers and 20 introns were extracted and aligned with MAFFT under default settings. Of the 110 automatically extracted intergenic spacers, one (*ndhA-ndhH*) was removed from further analysis owing to its length of only a single invariable nucleotide, whereas two (*rpl20-rps12* and *rps12-clpP*) were removed owing to their adjacency to gene *rps12*, which is likely to enforce idiosyncratic mutational dynamics given its discontinuous group II introns with complex secondary structures (Glanz and Kueck, 2009). Of the 20 automatically extracted introns, one (i.e. the intron of *trnK-TTT*) was removed from further analysis because it is the terminal region in the assembly of the large single-copy (LSC) in several of our samples and may, thus, exhibit an artificial assembly-induced length variability. The alignments of

the remaining 107 intergenic spacers and 19 introns were inspected for optimal positional homology and, where necessary, adjusted to improve alignment as discussed below.

Manual adjustment of sequence alignments

All alignment inspections were conducted by eye and all alignment adjustments by hand using PhyDE v.0.9971 (Müller et al., 2010). Both the inspections and the adjustments followed the alignment improvement rules described in Löhne and Borsch (2005), which are based on the observation that nucleotide substitutions and microstructural mutations do not typically occur at random but instead are often linked to structural or functional constraints and, thus, follow specific evolutionary mechanisms (Graham et al., 2000). For example, structural or functional constraints of plastome evolution can lead to a higher-than-expected frequency of short SSRs within plastid genomes, which should be accounted for during sequence alignment via a motif-based approach (Borsch and Quandt, 2009). However, many contemporary software algorithms for MSA have only a limited ability to identify conserved sequence motifs and to model length mutational events or other microstructural mutations. Small sequence inversions, for example, represent a particular challenge for most MSA software, as these inversions appear as a group of adjacent nucleotide substitutions to the alignment software unless manually re-inverted (e.g. Kim and Lee, 2005). Unadjusted sequence alignments of such inversions would lead to an overestimation of the implied amount of change between the sequences (Löhne and Borsch, 2005).

The aim of our alignment adjustments was to improve the positional homology among the aligned nucleotide sequences by applying motif-based adjustments on top of the automatic, software-driven alignments, especially for genome regions prone to microstructural mutations. For example, small sequence indels are length mutational events that typically correlate with other microstructural mutations (Kelchner, 2000; Morrison, 2006). Their exact alignment often requires an adjustment after the initial software-driven alignment: complete and uninterrupted indels need to be placed into the same columns, whereas overlapping indels need to be re-positioned to minimize the number of length mutational events (Löhne and Borsch, 2005). The sequence motif approach of alignment adjustment is, thus, predicated on hypothesized homologous sequence motifs. It aims to preserve these motifs but also acknowledges the occurrence of one or multiple overlapping length mutational events or microstructural mutations. We applied this approach and evaluated numerous cases of uncertain sequence alignments via visual inspections, followed by the manual, motif-based adjustment of the alignments wherever necessary. A summary of our full set of alignment adjustments is given in Table S1.

Several examples of our alignment adjustments are presented in Fig. 1 to illustrate both the alignment challenges involved and the solutions employed. Fig. 1a,c,d present examples of short- to medium-sized SSRs that were not recognized as sequence motifs by the alignment software and, thus, initially aligned with unrelated nucleotides: an SSR of 24 bp length in Fig. 1a (sequence motif “TTTCTACTTATACTACTAATATAA”), an SSR of 5 bp length in Fig. 1c (“GTGAT”), and an SSR of 5 bp length in Fig. 1d (“ATTATA”). The latter case is further complicated by the simultaneous presence of a preceding poly-T region in which the precise homology of the individual thymine nucleotides across the sequences cannot be readily determined, necessitating their exclusion from the sequence matrix. Manual alignment adjustments were conducted to improve the positional homology in each of these alignments. While it is true that the adjusted alignments could have theoretically been received through the application of specific gap penalties in the alignment software, the required penalty values would have been highly idiosyncratic per alignment, precluding the use of a global

gap penalty. The alignment adjustment displayed in Fig. 1b, by comparison, prevents the inclusion of numerous presumably incorrect SNPs between the in- and the outgroup, even though the total amount of length mutational events remains constant. The alignment adjustment displayed in Fig. 1e illustrates an unrecognized sequence inversion that is present in a subset of the aligned sequences (i.e. sequence motif “CCCCATCGG” in *Barclaya longifolia* KY284156, *Nymphaea alba* KU234277, *N. amazonum* MW057741, *N. lasiophylla* MW057731, *N. lotus* MW057729 and MW057730, *N. mexicana* MW057727, *Victoria amazonica* SmithEtAl2022, *V. boliviana* SmithEtAl2022 and *V. cruziana* KY001813 and SmithEtAl2022) and located in the centre of palindromic flanking sequences (i.e. “ACCT” and “TGGA,” respectively). By adjusting the local alignment, the inversion was inverted and re-integrated into the sequence matrix to prevent an overestimation of the implied amount of change between sequences; the inversion was additionally coded as a single mutational step in the accompanying indel matrix and, thus, integrated into the phylogeny reconstruction. Hypervariable sequence elements whose positional homology remained uncertain despite our best efforts in alignment adjustment (e.g. poly-A/T mini- and microsatellites) were excluded from the sequence matrices to prevent the inclusion of a false positive signal in the subsequent phylogeny reconstructions (Borsch and Quandt, 2009).

Coding of insertions/deletions and concatenation to phylogenetic matrix

In order to include the evolutionary signal of length mutational events and other microstructural mutations in our phylogenetic analyses, all indels and all sequence inversions larger than 3 bp were transliterated into presence-absence data. The indels were coded as binary characters in both the software-aligned and the manually adjusted alignments of the protein-coding genes, intergenic spacers and introns using the simple indel-coding scheme of Simmons and Ochoterena (2000) as implemented in 2matrix v.1.0 (Salinas and Little, 2014). Moreover, a total of 12 sequence inversions larger than 3 bp were identified during the alignment inspections (Table S1), coded as binary characters and added to the phylogenetic matrix. Such a procedure had been recommended by Kelchner and Wendel (1996) to account for the high levels of homoplasy exhibited by small sequence inversions in noncoding plastome regions. Consequently, the concatenated matrix of the 203 individual alignments represented a total of 145 609 characters before and 148 642 characters after indel coding and the addition of the sequence inversion characters.

Calculation of alignment and homoplasy statistics

In order to assess the impact of the motif-based alignment adjustments on alignment quality and the subsequent phylogenetic reconstructions, we calculated seven different summary statistics for each alignment before and after the alignment adjustments: (i) the length of each alignment, (ii) the number of gapped sites in each alignment (i.e. the number of sites that contain a gap in at least one of the aligned sequences), (iii) the number of polymorphic sites in each alignment (i.e. the number of sites that contain either a gap or a nucleotide substitution in at least one of the aligned sequences), (iv) the number of parsimony-informative sites (PIS) in each alignment, and (v) the consistency index (CI; Kluge and Farris, 1969) of each alignment, the rescaled consistency index (RC; Farris, 1989) of each alignment, and the retention index (RI; Farris, 1989) of each alignment. The last three statistics are homoplasy indices, and neighbour-joining trees inferred directly from the alignments were used for their calculation. All statistics were calculated in R (R Development Core Team, 2021) using the R packages ape v.5.2 (Paradis and Schliep, 2018) and phangorn v.2.4.0 (Schliep, 2011).

Inference of best-fitting nucleotide substitution models

In order to assess the impact of the motif-based alignment adjustments on the inference of best-fitting nucleotide substitution models, we applied ModelTest-NG v.0.2.0 (Darriba et al., 2019) on each of the 203 plastome regions to infer the best-fitting nucleotide substitution model before and after alignment adjustment. The Akaike information criterion (AIC; Akaike, 1974) was used as the statistical criterion for model selection, and models were considered distinct if they differed in the presence or absence of one or more model parameters.

Phylogenetic reconstruction

Phylogenetic tree reconstruction was conducted under the maximum parsimony (MP), the maximum-likelihood (ML) and the Bayesian inference (BI) criteria. Analyses via MP were conducted using TNT v.1.6 (Goloboff and Morales, 2023) with 10 replicates held per search step, the TBR branch-breaking algorithm, and 1000 equally parsimonious trees retained during the entire search; branch support under MP was calculated via 100 bootstrap (BS) replicates. Analyses via ML were conducted using RAxML v.8.2.9 (Stamatakis, 2014) under the thorough ML optimization option, with branch support calculated via 100 BS replicates using the rapid BS algorithm (Stamatakis et al., 2008). Analyses via BI were conducted with MrBayes v.3.2.5 (Ronquist and Huelsenbeck, 2003) under four parallel Markov chain Monte Carlo (MCMC) runs for a total of 50 million generations, with branch support given as posterior probability (PP) values. All BS values >95% and all PP values >0.95 are termed “near-maximum support” in this study. For BI, the sampling of independent generations and the convergence of the Markov chains were evaluated with Tracer v.1.7.1 (Rambaut et al., 2018); the initial 50% of all MCMC trees were discarded as burn-in, and the post-burn-in trees for each alignment were summarized as a 50% majority-rule consensus tree. Under both model-based inference criteria, the substitutions between nucleotide characters were modelled via the GTR + G + I nucleotide substitution model because the GTR model was found to be the best-fitting model across the majority of the individual and combined plastid sequence alignments under study. The substitutions within the presence/absence matrix of coded indels were modelled via the F81-like binary substitution model with a gamma-shaped rate variation of Lewis (2001). The plastome of *Cabomba caroliniana* was used as the outgroup in all phylogenetic reconstructions.

Evaluation of the impact of alignment adjustment and indel coding on tree inference

In order to evaluate the impact of motif-based alignment adjustments and the coding of indels on the results of phylogenetic tree inference, we tested the significance of the differences in the resulting tree topologies using a statistical framework. Specifically, we compared the likelihoods of competing tree topologies using the approximately unbiased (AU) test (Shimodaira, 2002) as implemented in CONSEL v.0.20 (Shimodaira and Hasegawa, 2001) under a significance threshold of $\alpha = 0.05$. The competing phylogenetic tree topologies were those inferred on the concatenated matrix of all coding and noncoding plastome regions under ML, either with or without alignment adjustments and either with or without the coding of indels. The nucleotide matrix for the AU tests (and, thus, the null hypothesis of the tests) represented the matrix with neither alignment adjustments nor indel coding. To better understand the importance of the AU test results for phylogenomic analysis, we also visualized the distribution of differences in PIS across all 203 plastome regions for the same comparisons (i.e. with and without the alignment adjustments, with and without coding of indels, and with and without both factors) and subdivided the results based on the statistical significance of the AU tests.

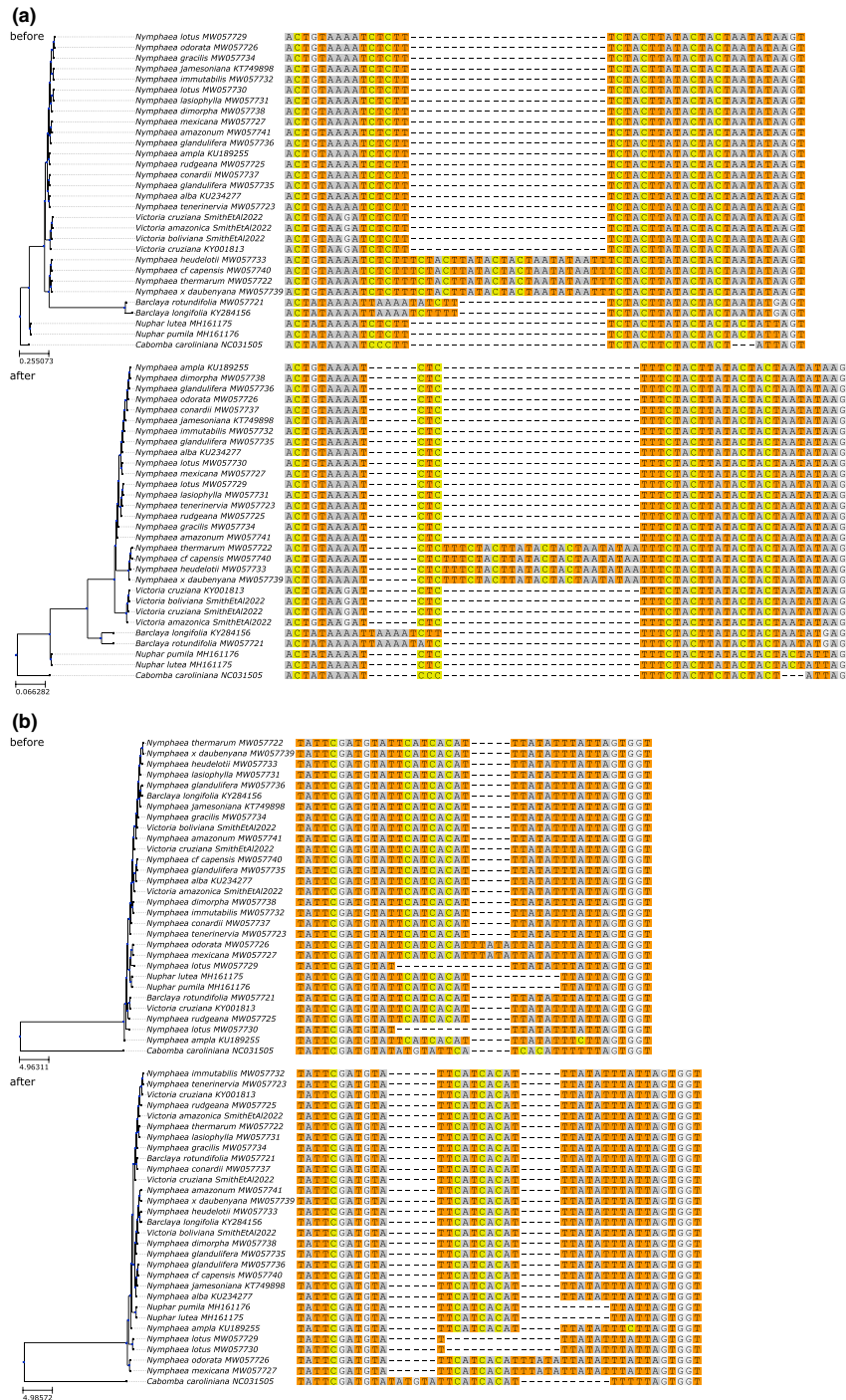
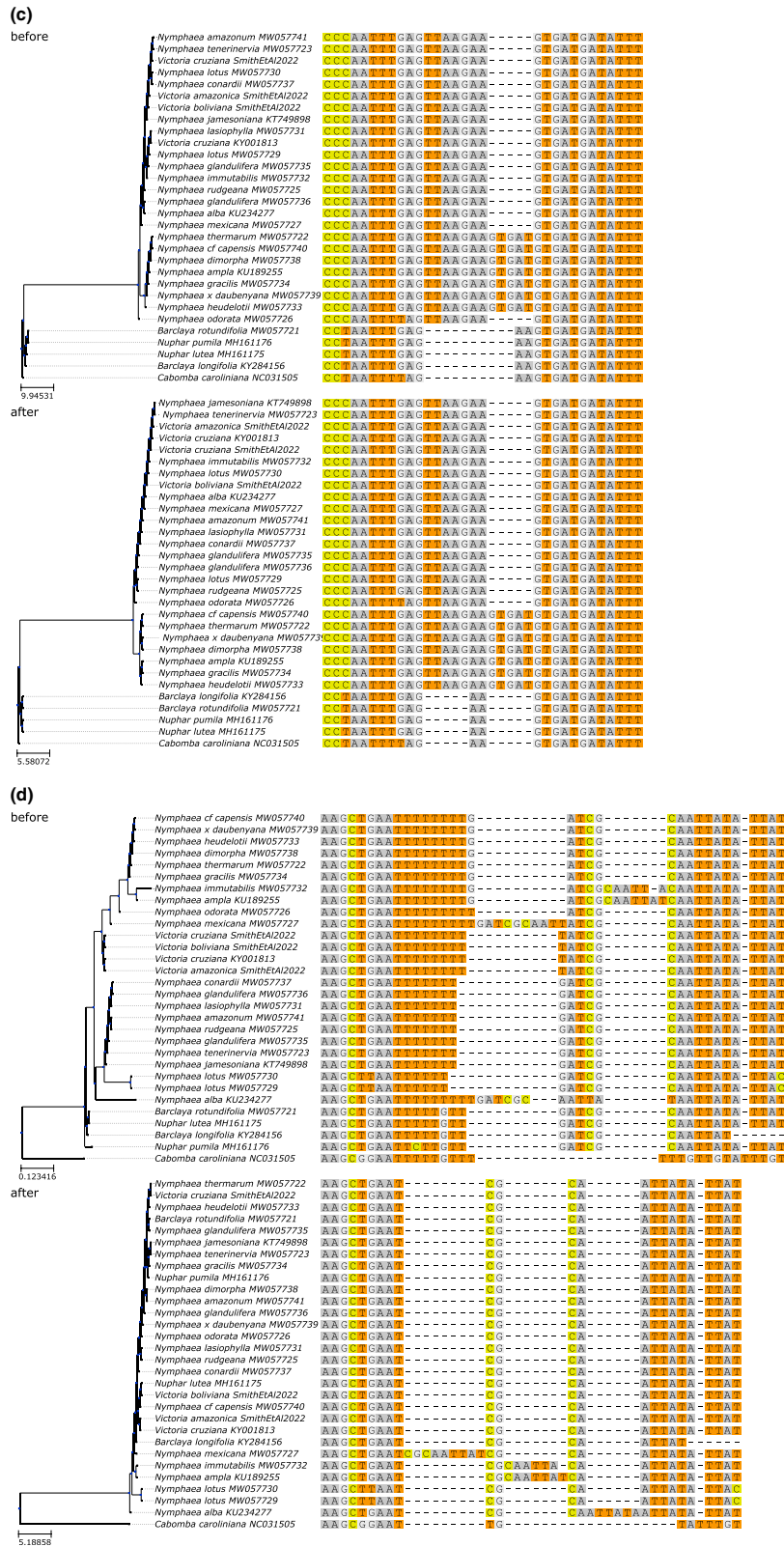


Fig. 1. Examples of motif-based alignment adjustments conducted in this investigation and their impact on phylogenetic tree inference. Displayed are sections of the DNA sequence alignments of (a) the gene *ycfI* (positions 4279–4341 of its unadjusted alignment), (b) the first intron of gene *clpP* (positions 759–801 of its unadjusted alignment), (c) the intergenic spacer between genes *atpH* and *atpI* (positions 253–287 of its unadjusted alignment), (d) the intergenic spacer between genes *petA* and *psbJ* (positions 40–90 of its unadjusted alignment), and (e) the intergenic spacer between genes *psbT* and *psbN* (positions 17–62 of its unadjusted alignment) of the plastid genomes. For each example, the status before (top) and after (bottom) the motif-based alignment adjustment is displayed. Each alignment is preceded by a phylogenetic tree that was inferred through a partitioned ML analysis under the use of the GTR nucleotide substitution model and the coding of indels.



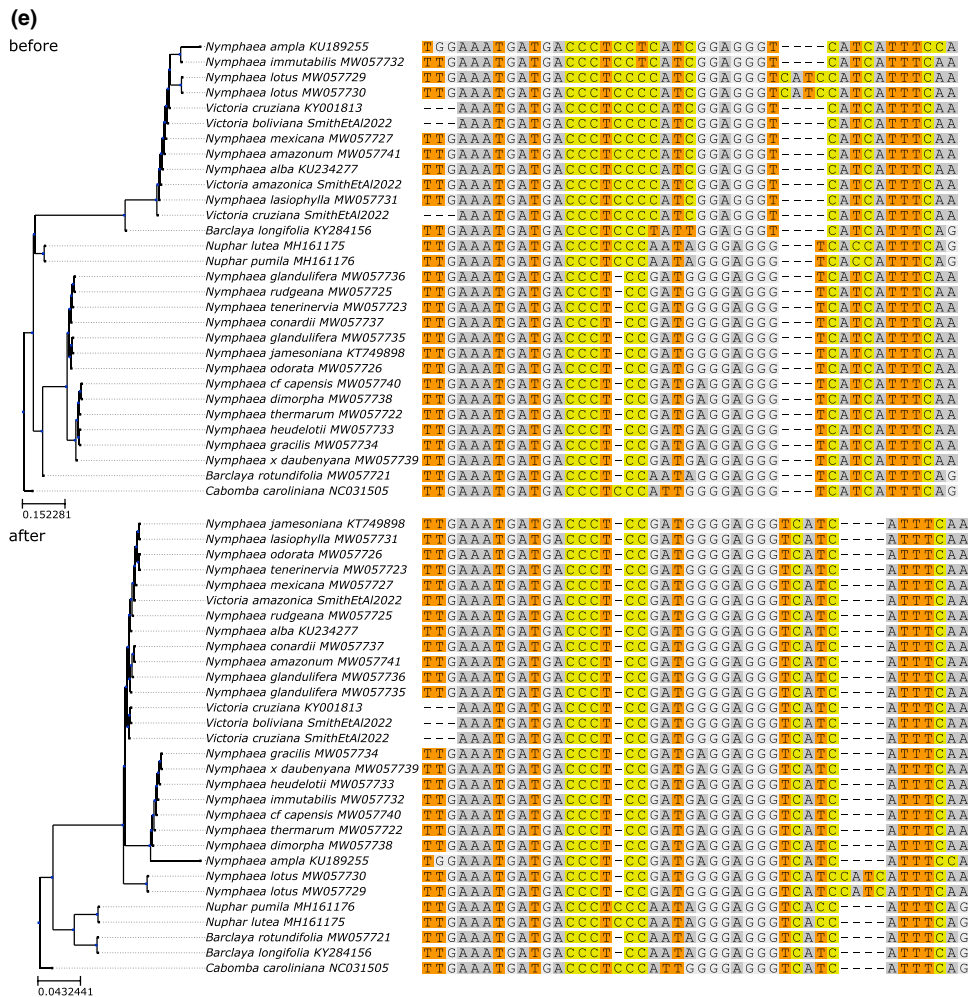


Fig. 1. Continued

Ancestral character state reconstructions of sequence inversions

A total of 12 sequence inversions larger than 3 bp were identified across the analysed plastid genomes (Table S1). Next to reverse-complementing and then re-integrating these inversions into the alignments as part of our alignment adjustments, we also encoded them as presence-absence data into the phylogenetic matrix (see above) and assessed if their occurrence matched any of the speciation events implied by our resulting phylogeny. Specifically, we inferred the ancestral character states of each of these sequence inversions using the marginal ancestral state estimation method of Yang et al. (1995) on the phylogenetic tree with the best likelihood score as reconstructed from the concatenation of the adjusted alignments.

Results

Genome structure and gene content

All newly generated plastomes exhibited the typical quadripartite genome structure of land plants

(Mower and Vickrey, 2018): they comprised one LSC and one small single-copy (SSC) region, separated by two identical inverted repeats (IRs). The total length of the newly generated genomes varied between 160 043 bp (*Nuphar lutea*) and 158 288 bp (*Nymphaea tenerinervia*), with similar length variations across the four regions of the plastid genome. Gene content was found to be highly conserved: all newly generated genomes exhibited a total of 79 different protein-coding genes (eight duplicated in the IR), 30 tRNA genes (seven duplicated in the IR), and four rRNA genes (all duplicated in the IR). Ten of the coding regions contained one (*atpF*, *ndhA*, *ndhB*, *petB*, *petD*, *rp12*, *rpoC1* and *rps16*) or two (*clpP* and *yef3*) introns across all samples. The plastid genomes of *N. thermarum* and *N. heudelotii* were found to be identical in both length and sequence. *Nymphaea immutabilis* (MW057732) exhibited the longest plastome sequence of the core Nymphaeaceae, and a circular representation of this

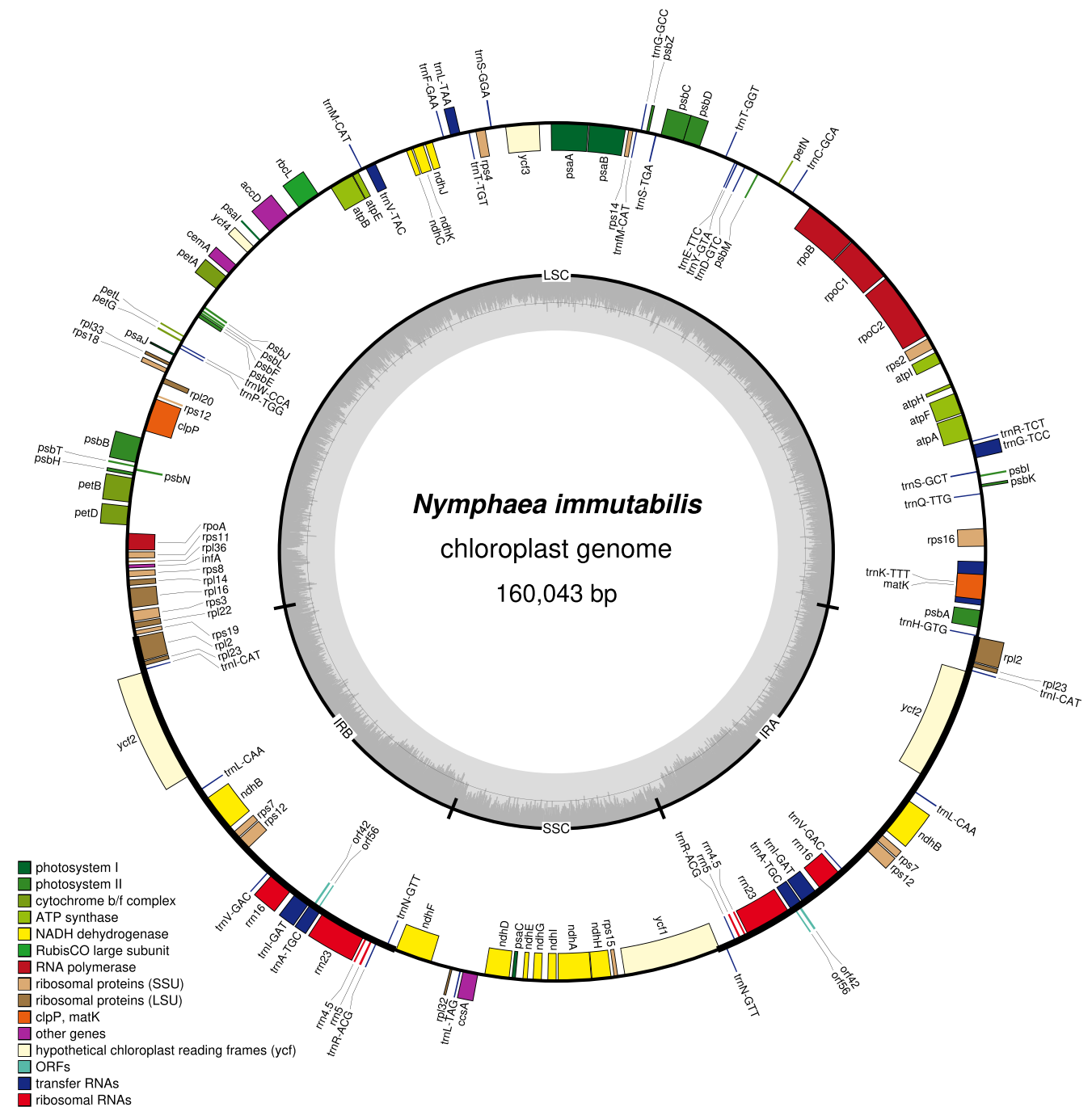


Fig. 2. Circular representation of the plastid genome of *Nymphaea immutabilis* (MW057732). Genes displayed on the outside of the outer circle have a clockwise transcription, those on the inside a counter-clockwise transcription. The inner circle displays the boundaries of the LSC, SSC and IR regions, as well as the GC content across the genome.

plastome is displayed *pars pro toto* for the plastomes of the other taxa (Fig. 2). The concatenation of all 203 coding and noncoding plastome regions resulted in a dataset of 2996 PIS before and 2851 PIS after alignment adjustment, with the proportion of gaps or undetermined characters <1% in either case.

Effect of alignment adjustment on alignment and homoplasy statistics

The motif-based adjustments of the individual DNA sequence alignments were found to have a considerable effect on most of the alignment summary statistics: on

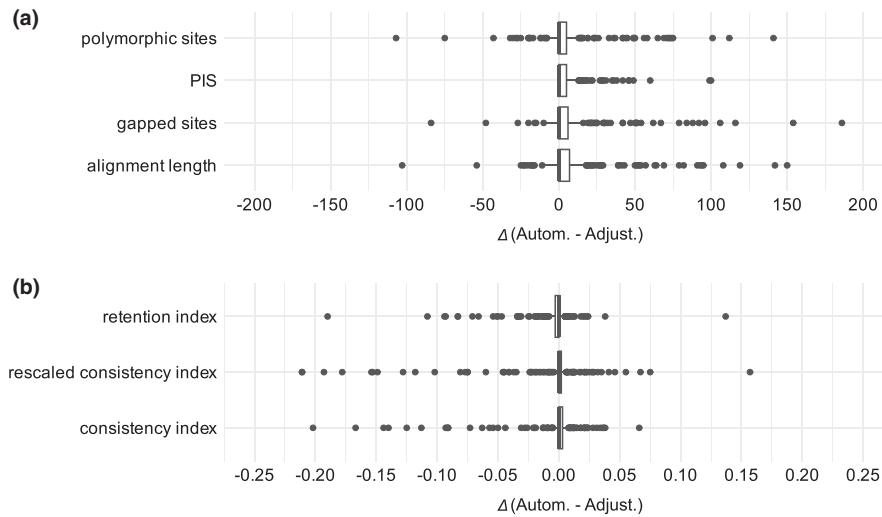


Fig. 3. Distributions of the differences before and after alignment adjustment in (a) alignment summary and (b) homoplasy statistics calculated across all 203 plastome regions under study. The distributions are visualized as horizontal boxplots; outlier values are not displayed. Autom., automatically generated alignment (i.e. before alignment adjustment); Adjust., after alignment adjustment; PIS, parsimony-informative sites.

average, the alignment adjustments resulted in lower homoplasy levels as well as lower summary statistic values across most of the plastome loci under study (Fig. 3). Specifically, alignment length, the number of gapped sites per alignment, the number of polymorphic sites per alignment and the number of PIS per alignment were, on average, each found to be lower after than before the alignment adjustments (Fig. 3a). By comparison, each of the homoplasy statistics was, on average, found to be higher after than before the alignment adjustments. Because homoplasy statistics are negatively correlated with the level of homoplasy in an alignment, an increase in the homoplasy statistic represents a reduction in the actual homoplasy level (Fig. 3b). Specifically, our alignment adjustments consistently reduced the level of homoplasy in the sequence matrices, even if it occasionally also reduced the number of PIS per alignment. Each of the 203 individual alignments as well as all concatenated alignments, both before and after alignment adjustment and both with and without the coded indel characters, are available on Zenodo at <https://doi.org/10.5281/zenodo.7860937>.

Effect of alignment adjustment on the inference of nucleotide substitution models

Our motif-based alignment adjustments were found to have a largely idiosyncratic impact on the inference of best-fitting nucleotide substitution models. We found that only three coding regions (i.e. 4% of them), but 31 intergenic spacer regions (29% of them) and four introns (21% of them) exhibited different best-fitting nucleotide substitution models before and after alignment adjustment (Tables S2–S4). These

differences primarily reflected the presence or absence of auxiliary model parameters (i.e. the gamma-distributed rate variation among sites and the number of invariant sites) rather than a difference in the actual number of substitution rates. Moreover, the set of plastome regions exhibiting different best-fitting nucleotide substitution models before and after alignment adjustment did not match the set of regions with reduced levels of homoplasy or with significantly different tree topologies before and after alignment adjustment. Additionally, differences in log-likelihoods of model fit before and after alignment adjustment did not coincide with differences in best-fitting substitution models, indicating that our motif-based alignment adjustments were only one of multiple factors affecting the inference of best-fitting nucleotide substitution models.

Effect of alignment adjustment and indel coding on tree inference

In order to evaluate if motif-based alignment adjustment influences not only alignment summary and homoplasy statistics but also phylogenetic tree inference, we tested the significance in the difference of phylogenetic tree topologies reconstructed under ML using the AU test. The same test was also employed to assess the effect of indel coding on phylogenetic tree inference. Significant differences in the phylogenetic reconstructions were detected as a result of alignment adjustments (Table 2): we found that the topologies of the best ML trees inferred before and after the alignment adjustments were significantly different, but that this was only the case when the indels remained uncoded. Conversely, the topologies of the best ML trees inferred with and

Table 2
Statistical comparison of competing phylogenetic tree topologies as inferred through AU tests on the concatenated plastome matrix

Constraint	Visualization of topology	AU test <i>P</i> -value
Before alignment adjustment, WOC (positive control)	Fig. 5a	0.095
After alignment adjustment, WOC	Fig. 5b	0.002*
Before alignment adjustment, SIC	Fig. 6a	0.896
After alignment adjustment, SIC	Fig. 6b	0.093

The topological constraints represent the topologies of the best ML trees inferred with or without alignment adjustments and with or without the coding of indels. Significant *P*-values are indicated by an asterisk.

without indel encodings were not significantly different unless alignment adjustments were present. In summary, phylogenetic trees inferred from complete plastome sequence data showed significant differences in their topology regarding the application of motif-based alignment adjustments, but only in the absence of indel encodings. This finding indicates that the coding of indels may have an equally large effect on phylogenetic signal as the alignment adjustments.

A somewhat different pattern of significant differences between tree topologies was detected when the aligned genes, intergenic spacers and introns were evaluated individually (Tables S5–S7): depending on the plastome region under study, we found significant differences in tree topology when comparing alignment

adjustment, indel coding and both factors combined. For example, multiple significant AU tests were recorded (i.e. for nine genes, 21 intergenic spacers and five introns) when the compared topologies differed owing to the presence or absence of alignment adjustments and the indels were coded.

In order to better understand the effect of motif-based alignment adjustment and indel coding on the observed differences in tree topology, we visualized the differences in PIS for the same comparisons as evaluated by the AU tests and then subdivided the distributions by AU test outcome. The resulting distributions showed that the differences in PIS were similar between genomic regions with significant AU test outcome and those without (Fig. 4). Although the difference distributions illustrating the effect of indel coding (Fig. 4a) or alignment adjustment (Fig. 4b) on tree topology were similar across AU test outcomes, the distribution illustrating the effect of both factors combined was not (Fig. 4c): the combined comparison indicated a stronger variation in the number of PIS among plastome regions with significant AU test outcome than those without. A significant AU test outcome for individual genome regions is, thus, likely to be the result of a change in the phylogenetic signal caused by both factors simultaneously (i.e. alignment adjustment and indel coding) than by one factor alone.

Phylogenetic relationships

Our phylogenetic reconstructions resulted in highly resolved phylogenetic trees with high levels of node

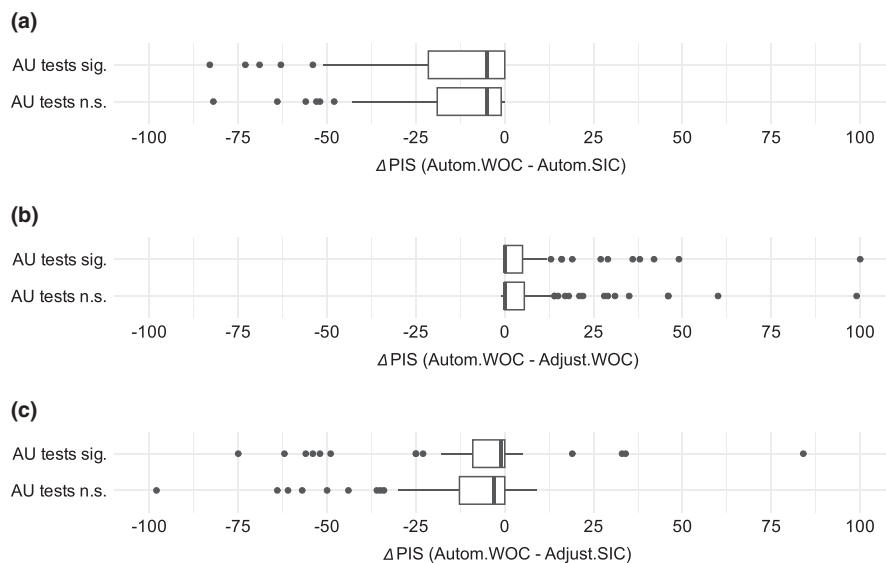


Fig. 4. Distributions of the differences before and after alignment adjustment in parsimony-informative sites between (a) the automatically generated sequence alignments either with (“SIC”) or without (“WOC”) indel coding, (b) the automatically generated and the adjusted sequence alignments, both without indel coding, and (c) the automatically generated and the adjusted sequence alignments, with only the latter containing indel encodings. Each difference distribution is subdivided into cases with significant AU test outcomes (“sig.”) and those without (“n.s.”). All distributions are visualized as horizontal boxplots. Abbreviations are as in Fig. 3.

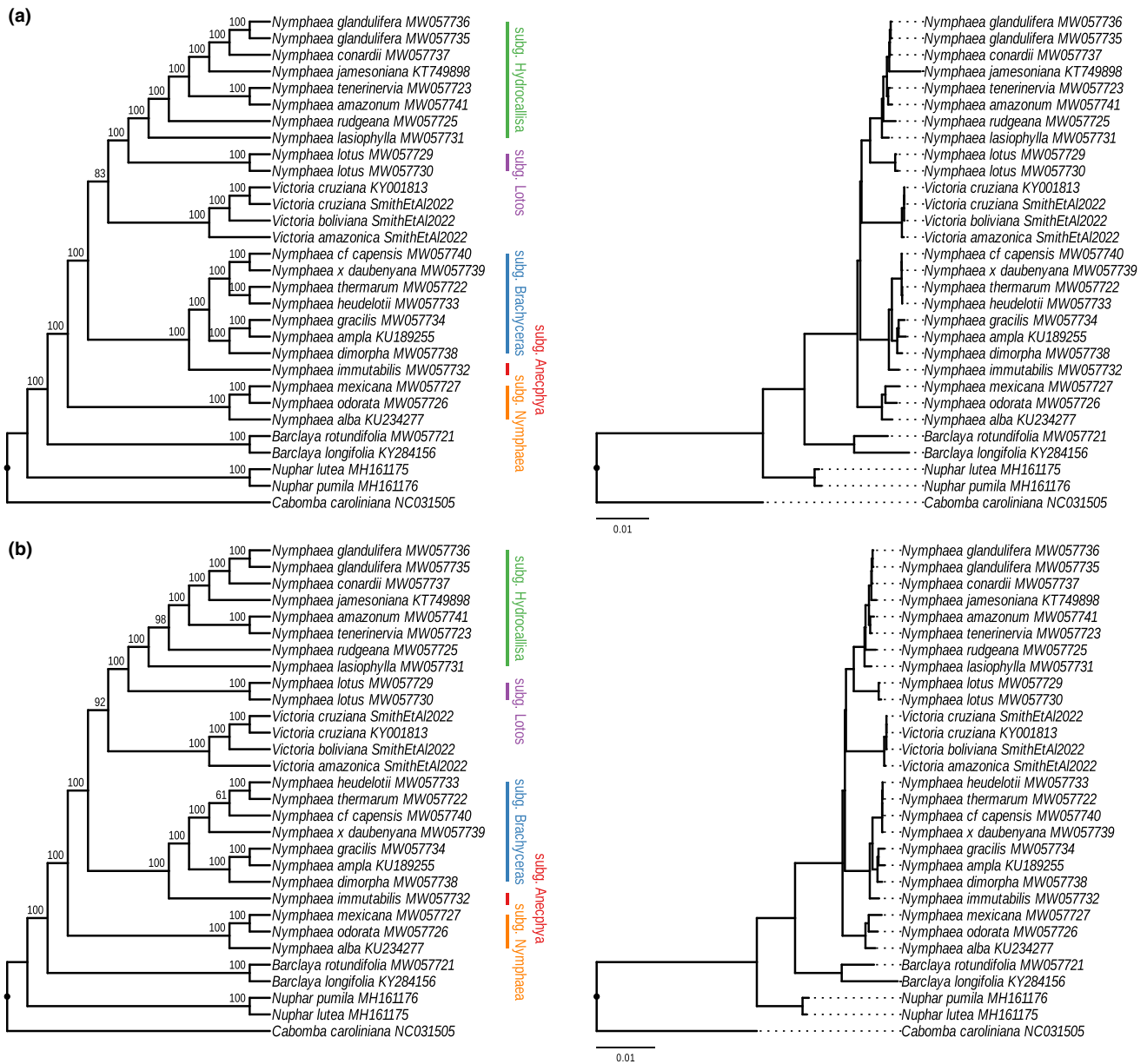


Fig. 5. Phylogenetic relationships of Nymphaeaceae inferred under ML on the concatenated data set of all coding and noncoding plastome regions when indels are uncoded. The trees displayed represent those with the best likelihood score inferred either (a) before or (b) after motif-based alignment adjustment and are visualized as cladograms with statistical node support (left) and corresponding phylograms with exact branch lengths (right). Bootstrap support >50% is given above the branches of each cladogram. All trees were rooted using *Cabomba caroliniana* as outgroup.

support for almost all clades. The reconstructions under ML, for example, inferred relationships among the core Nymphaeaceae that were highly congruent with previous molecular phylogenetic studies and exhibited high bootstrap support for almost every node, irrespective of alignment adjustment or indel coding (Figs 5 and 6). Each of the five subgenera of *Nymphaea* was recovered as monophyletic with maximum BS support. The inferences of subgenus *Nymphaea* as sister to the other subgenera of *Nymphaea*,

the sister relationship between subgenera *Brachyceras* and *Anecephya*, and the sister relationship between subgenera *Hydrocallis* and *Lotos* were also received with maximum support. Likewise, the clade formed by subgenera *Brachyceras* and *Anecephya* as sister to the clade that comprises subgenera *Hydrocallis* and *Lotos* as well as genus *Victoria* was fully supported. Only the sister relationship between *Victoria* and subgenera *Hydrocallis* and *Lotos* was supported by less than full support, depending on alignment adjustments and the

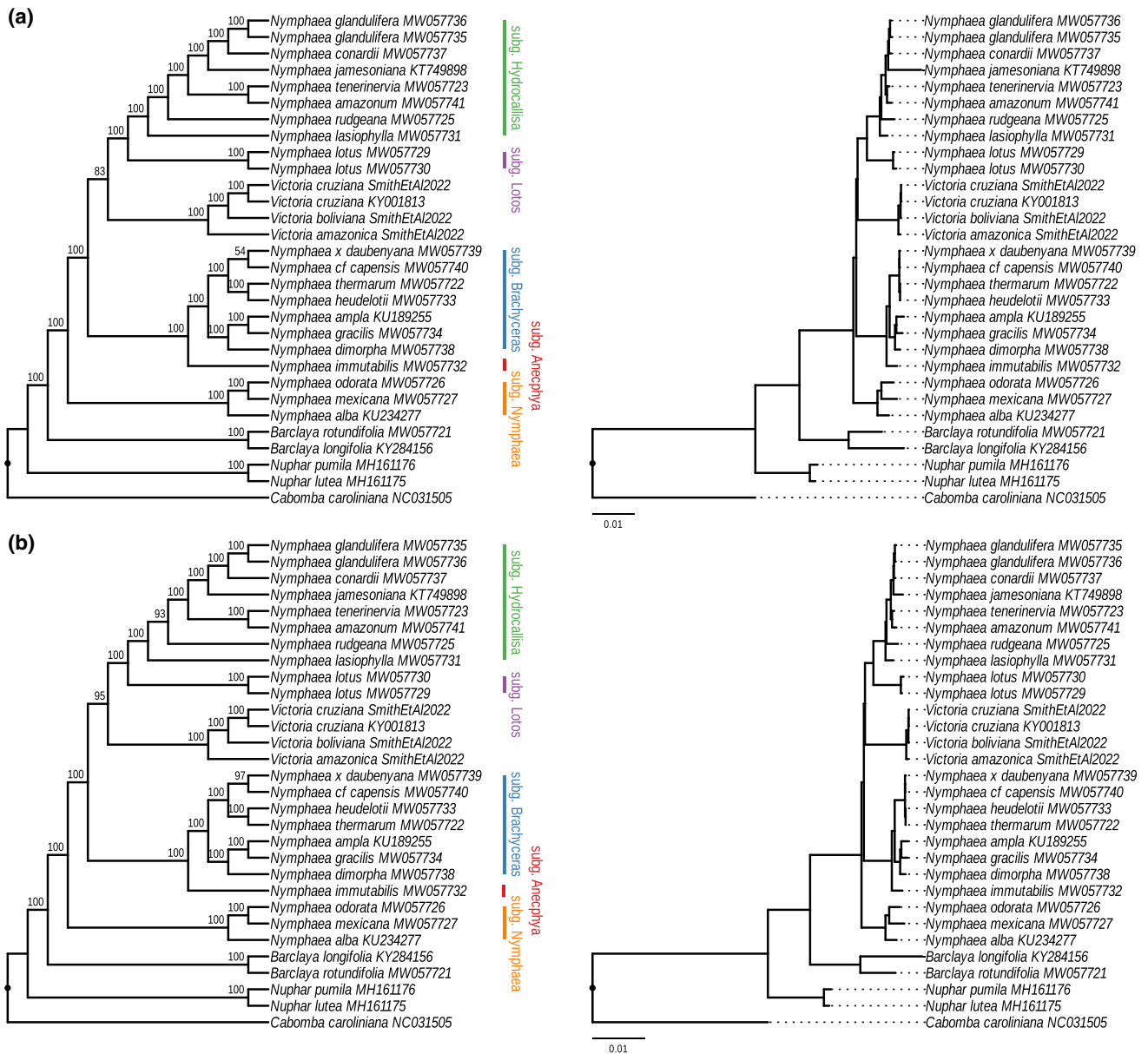


Fig. 6. Phylogenetic relationships of Nymphaeaceae inferred under ML on the concatenated data set of all coding and noncoding plastome regions when indels are coded. The trees displayed represent those with the best likelihood score inferred either (a) before or (b) after motif-based alignment adjustment and are visualized as cladograms with statistical node support (left) and corresponding phylograms with exact branch lengths (right). All settings are identical to the ML tree inference without the coding of indels (Fig. 5).

coding of indels: the relationship was supported by BS 83% without adjustments and without indel coding (Fig. 5a), by BS 92% with adjustments but without indel coding (Fig. 5b), by BS 83% without adjustments but with indel coding (Fig. 6a), and by BS 95% with both adjustments and indel coding (Fig. 6b). The only difference in the phylogenetic reconstructions pertaining to alignment adjustment was found in the relationships among the species of subgenera *Brachyceras*: the reconstruction with alignment adjustment but without indel coding (Fig. 5b) did not support a sister

relationship between *Nymphaea cf. capensis* and *Nymphaea x daubenyana*, whereas all other reconstructions did (Figs 5a and 6).

The phylogenetic reconstructions under MP and BI mirrored those conducted under ML. The relationships retrieved between the five subgenera of *Nymphaea* were identical to those inferred via ML and typically exhibited maximum BS or PP support (Figs S1–S4). The inferred 50% majority-rule consensus trees differed in their topology regarding only one relationship: a sister relationship between *Nymphaea cf. capensis*

and *Nymphaea* × *daubenyana* among the species of subgenera *Brachyceras* was either not supported (i.e. when alignments were adjusted and indels were coded; Fig. S1B) or not resolved (Fig. S4B), whereas all other MP and BI reconstructions did support this relationship (Figs S1a, S2, S3 and S4A).

Ancestral character state reconstructions of selected sequence inversions

The reconstruction of the ancestral character states of the 12 plastome sequence inversions indicated that only a few were synapomorphies for lineages within Nymphaeaceae (Fig. 7). For example, inversion 1 is a synapomorphy of the core Nymphaeaceae plus *Barclaya*, whereas inversions 9 and 11 are synapomorphies of the three species of *Victoria*. Inversion 11 was found within gene *atpE*, whereas most other inversions (i.e., ten of 12) were found within intergenic spacers. The level of homoplasy differed among the inversions: character state changes were most often observed in the *psbT-psbN* spacer (inversion 6) and the *trnH-psbA* spacer (inversion 10), with both spacers exhibiting multiple gains and losses of the inversions.

Discussion

Phylogenetic relationships among the core Nymphaeaceae

Our phylogenetic reconstructions recovered multiple strongly supported relationships among the core Nymphaeaceae (Figs 5, 6 and S1–S4), of which at least four are congruent with the results of previous investigations. First, our results are congruent with the observation by Borsch et al. (2007) that the core Nymphaeaceae exhibit four distinct clades: one clade formed by *Euryale* and *Victoria*, and three clades formed by the different subgenera of *Nymphaea*. The clades formed by *Nymphaea* are a clade of the autonymic subgenus (i.e. subgenus *Nymphaea*), a clade comprising subgenera *Anecphyta* and *Brachyceras*, and a clade comprising subgenera *Hydrocallis* and *Lotos*. Second, our results corroborate the position of the temperate subgenus *Nymphaea* as the earliest-diverging lineage in the core Nymphaeaceae (BS 100/PP 1.0 across all analyses). This position had initially been reported by Löhne et al. (2007) with moderate support in their analysis of a matrix of eight noncoding plastid regions plus *matK*. By comparison, Borsch et al. (2007) recovered a topology inconsistent with these results, where the *Euryale-Victoria* clade was resolved as sister to a poorly supported monophyletic genus *Nymphaea*, with subgenus *Nymphaea* recovered as the first diverging branch of that lineage. A similar

phylogenetic position of subgenus *Nymphaea* was recovered by nrITS sequence data and a parsimony analysis of a matrix of 62 morphological and anatomical characters (Borsch et al., 2008). Third, our results confirm the sister relationship between the subgenera *Anecphyta* and *Brachyceras* of *Nymphaea*, which had initially been reported by Löhne et al. (2007). Whereas the Australian subgenus *Anecphyta* was found to be monophyletic across all analyses and genomic partitions in previous studies (e.g. Borsch et al., 2007; Löhne et al., 2007, 2008a), the status of the pantropical subgenus *Brachyceras* remained largely unresolved. Borsch et al. (2007) recovered several species of subgenus *Brachyceras* in a polytomy with a clade comprising species of subgenus *Anecphyta* using plastid *trnT-trnF* sequence data, while a subsequent analysis with a more comprehensive species sampling provided similarly weak evidence for the monophyly of this clade under plastid genome data (Borsch et al., 2011). The present investigation provides robust evidence for the monophyly of subgenus *Brachyceras* from the plastid genome. Future studies should, however, include *Nymphaea petersiana*, which is native to the Malawi lake region in Africa and was previously recovered in subgenus *Lotos* instead of subgenus *Brachyceras* (Borsch et al., 2007, 2008). This placement indicated that the broad circumscription of *Nymphaea nouchali* of subgenus *Brachyceras* by Polhill and Verdcourt (1989), who subsumed several other species, including *N. petersiana*, as infraspecific entities, was highly unnatural. Unfortunately, the DNA quality of our herbarium specimens of *N. petersiana* was insufficient for an inclusion in this study. Fourth, our results are congruent with the reports of Löhne et al. (2007), who suggested a close relationship of *Victoria* to subgenera *Hydrocallis* and *Lotos* of *Nymphaea* and, by extension, the paraphyly of genus *Nymphaea*. Our reconstructions recovered the relationship between *Victoria* and the subgenera *Hydrocallis* and *Lotos* with high node support, but only when alignments had been adjusted; in the absence of such adjustments, the relationship was supported with a reduced level of node support (i.e. BS 83 ML; Fig. 6a). The phylogenetic position of *Victoria* in the same clade as species of *Nymphaea* subgenera *Hydrocallis* and *Lotos* is also supported by the floral biology of these species, as all three lineages exhibit night blooming.

Phylogenetic relationships within subg. Hydrocallis

Our phylogenetic reconstructions identified relationships within genus *Nymphaea* that had previously been hypothesized but remained untested. Specifically, our reconstructions resolved some phylogenetic relationships within subgenus *Hydrocallis* with high node support (Figs 5 and 6). Previous phylogenetic

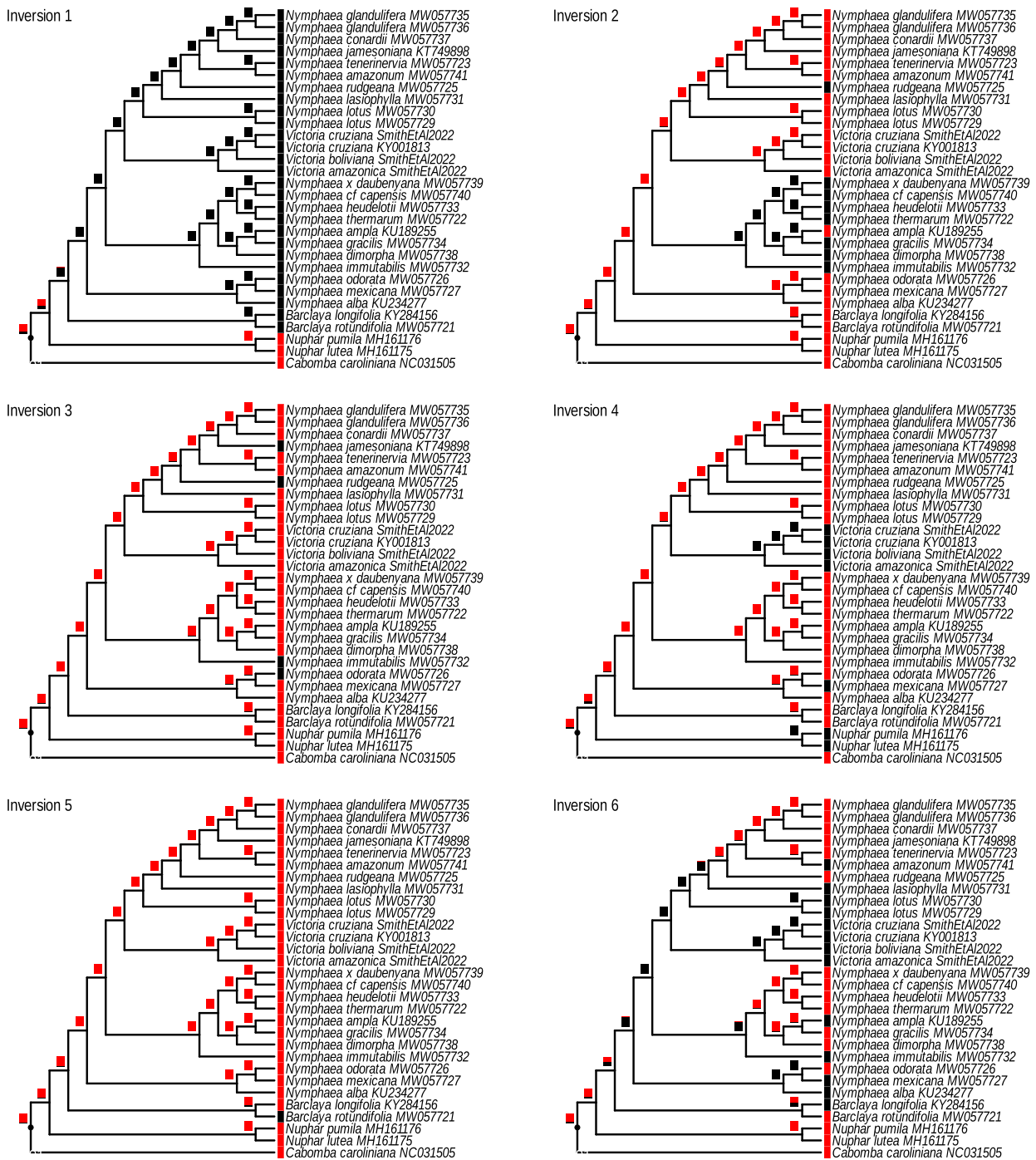


Fig. 7. Ancestral and contemporary states of 11 small sequence inversions that were found in the plastomes of Nymphaeaceae during the alignment inspections. The presence of an inversion in the plastomes of specific species is indicated by a black bar, the absence by a red bar. The inversions are located in the following plastome regions: inversion 1 in the *ndhC-trnV* intergenic spacer, inversion 2 in the *petA-psbJ* intergenic spacer, inversion 3 in the *petD-rpoA* intergenic spacer, inversion 4 in the *psbL-trnS-TGA* intergenic spacer, inversion 5 in the *psbE-petL* intergenic spacer, inversion 6 in the *psbT-psbN* intergenic spacer, inversion 7 in the *psbZ-trnG-GCC* intergenic spacer, inversion 8 in the *rps18-rpl20* intergenic spacer, inversion 9 in the *trnF-GAA-ndhJ* intergenic spacer, inversion 10 in the *trnH-psbA* intergenic spacer and inversion 11 in gene *atpE*. Another inversion (inversion 12) is located in the intron of *rpl16* but has the same distribution as inversion 5 and is, thus, not visualized here.

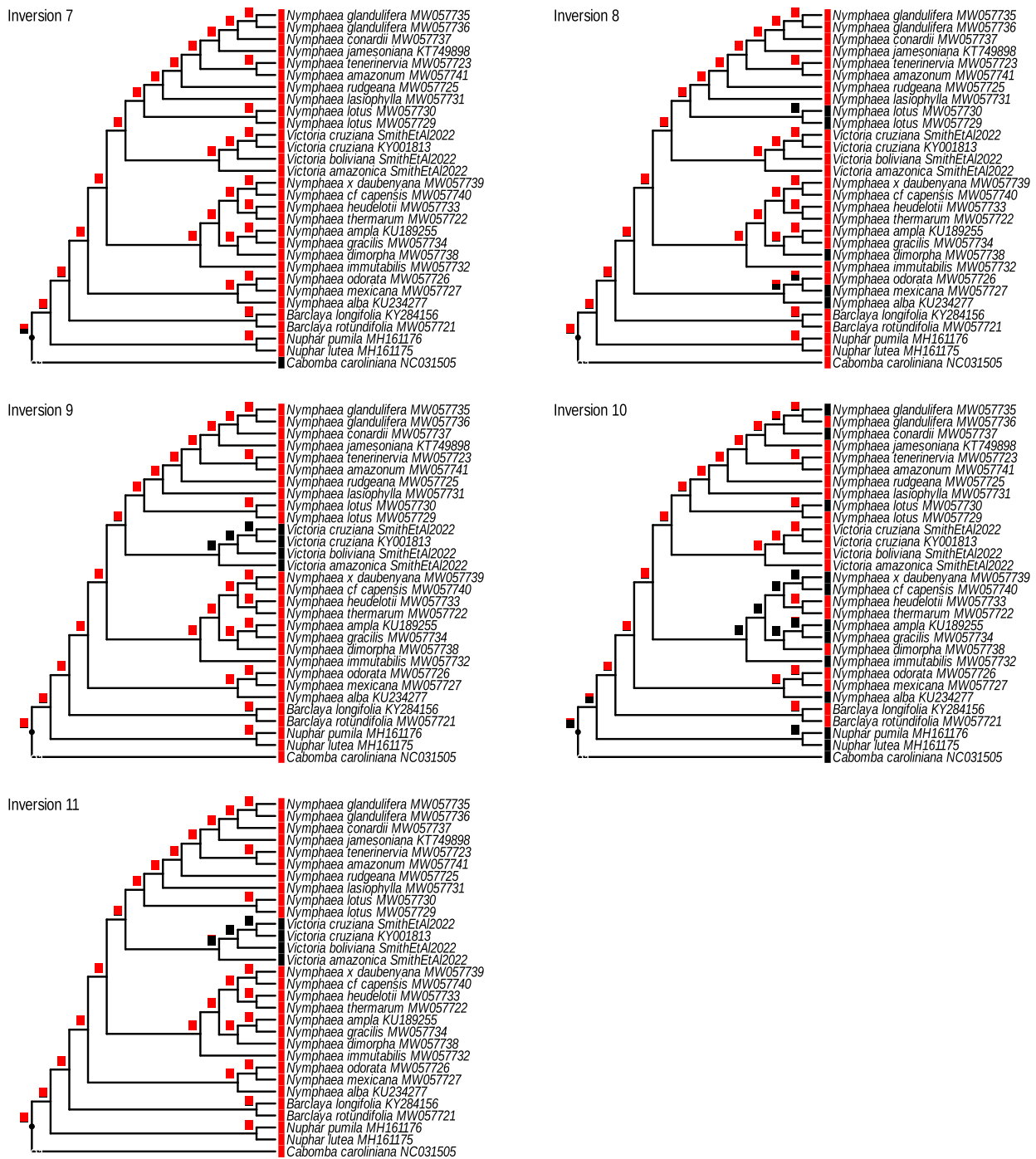


Fig. 7. Continued

reconstructions of subgenus *Hydrocallis* were limited to a suboptimal character base and unable to fully resolve relationships within this shallow clade. Borsch et al. (2007) found that resolution among members of this subgenus was supported by the sequence variation of the AT-rich elements of the P8 loop of the *trnL*

intron, which was only the case for this subgenus. The same phylogenetic signal also indicated a close relationship between *N. amazonum* and *N. tenerinervia*, which is supported with full node support by our results as well. Furthermore, our reconstructions provided further corroboration for a close relationship

between *N. conardii*, *N. glandulifera* and *N. jamesoniana* of subgenus *Hydrocallis*, which had been suggested by Borsch et al. (2007) but not yet affirmed. Each of these findings are in line with a crown-group age of subgenus *Hydrocallis* of approximately 7 Ma, which had been suggested by Löhne et al. (2008b) and renders this subgenus the most recently diverged among the five subgenera of *Nymphaea*.

Wiersema (1987) suggested that *N. lasiophylla* and *N. lingulata* might represent an ancestral lineage within subgenus *Hydrocallis*, given that these species share several morphological characters with other subgenera of *Nymphaea* (e.g. tapering capillary appendages, a less ordered arrangement of the perianth, an absence of petaloid stamens and a presence of 3'-methylated flavonoids). *Nymphaea lasiophylla* was not included in any previous molecular phylogenetic analysis, rendering this question unresolved. The results of our phylogenetic reconstructions support the suggestion by Wiersema (1987), as *N. lasiophylla* is identified as sister to all other members of subgenus *Hydrocallis* with maximum node support (Figs 5 and 6). The phylogenetic placement of *N. lingulata*, which was proposed as part of a lineage together with *N. oxypetala* and *N. rudgeana* (Borsch et al., 2007), remains unresolved, as no molecular phylogenetic analysis had so far evaluated this question. Our reconstructions partially agree with the suggestion by Borsch et al. (2007) regarding *N. rudgeana*, as that species is resolved as the second branch within subgenus *Hydrocallis*. The unusually high chromosome number of *N. rudgeana* ($2n = 42$) led Wiersema (1987) to speculate whether that species may be the result of interspecific hybridization, potentially involving a member of subgenus *Lotos*. Such an evolutionary reticulation would be consistent with the similarity in several leaf characters between *N. rudgeana* and *N. lotus*. However, our results provide support for *N. rudgeana* as a member of the crown group of subgenus *Hydrocallis* given its distinct plastome sequence. Such a placement was reported by Borsch et al. (2014) based on nrITS sequence data, although a strong uniparental bias during concerted evolution toward the dominant ribotype of subgenus *Hydrocallis* could also have caused such a result, assuming that a paternal ancestor from a different sub-lineage was involved in forming this species.

Our phylogenetic reconstructions support the relationship of *N. conardii*, *N. glandulifera* and *N. jamesoniana* as successive sister species within subgenus *Hydrocallis* with high confidence, which aligns with morphological, ecological and cytological evidence on these species. Wiersema (1987) suggested that the predominantly, if not exclusively, autogamous species *N. jamesoniana* is likely to be related to *N. gardneriana* and *N. conardii*, but that it stands out by the smallest seeds of any *Nymphaea* species and a prominent

pattern of reticulate venation on the lower leaf surface. The phylogenetic relationships identified in this study agree with the observation that *N. conardii*, *N. jamesoniana* and several other related species exhibit a distinct chromosome number of $2n = 28$, but also allow one of the two scenarios of chromosome evolution deemed less likely by Wiersema (1987): a scenario in which the chromosome number of *N. conardii*, *N. jamesoniana* and relatives is caused by a series of aneuploidy increases from an ancestral state of $2n = 18$. The congruence between the relationships proposed by Wiersema (1987) and those identified here are likely to be a consequence of the dense species sampling achieved in this investigation, as no fewer than half of all the species accepted by Wiersema (1987) were included in our phylogenetic reconstructions. However, several new taxa have since been described from Brazil based on a partially distinct morphology and assigned to *Nymphaea* subgenus *Hydrocallis* (de Lima and Guilietti, 2013; de Lima et al., 2021); future phylogenetic studies should include these potentially new species as well to better understand the evolutionary history of subgenus *Hydrocallis*.

Phylogenetic relationships within subgenus Brachyceras

Our phylogenetic reconstructions also resolved the relationships within the pantropical subg. *Brachyceras* with high node support (Figs 5 and 6 and S1–S4). The study of Löhne et al. (2007) had included only two taxa of this subgenus (i.e. the Mexican species *N. gracilis* and the African species *N. micrantha*), which were recovered as sister species, whereas the more comprehensive study of Borsch et al. (2011) had recovered a largely unresolved clade of species from the Americas as part of a polytomy comprising several lineages of African species of *Nymphaea*. Furthermore, nuclear ribosomal sequence data had indicated that the American species were nested among several palaeotropical lineages of *Nymphaea*, suggesting a long-distance dispersal out of Africa for the origin of the Neotropical species of *Nymphaea*, which was estimated to have occurred in the late Miocene (~ 10 Ma; Borsch et al., 2011). Our phylogenetic reconstructions are more resolved. First, a sister relationship of the Madagascar endemic *N. dimorpha* (previously known as *N. minuta*) to a clade comprising the Neotropical species *N. ampla* and *N. gracilis* was recovered with full node support. This relationship is surprising, as the sequence data of the nrITS had suggested a sister relationship of *N. dimorpha* to the West African species *N. guineensis* (not sampled here), which together had been found sister to all other species of subgenus *Brachyceras* (Borsch et al., 2011). Second, a close relationship between the African species *N. thermarum*, *N. heudelotii* and *N. capensis* was identified. Our

sample of *N. thermarum* is an F₁ individual of the type-material from hot springs in Rwanda (Fischer, 1988) that was cultivated through self-pollination and can, thus, be considered a part of the original collection. We found that its plastome sequence was identical to that of *N. heudelotii*, which is a larger, morphologically distinct species with purple flowers that also occurs in Rwanda. This observation could be explained through a recent plastid capture event—a phenomenon that has been reported from multiple plant groups (e.g. Liu et al., 2020; Baldwin et al., 2023)—or, alternatively, by a very recent species divergence. Future investigations should test for both, evidence of a plastid capture event as well as the possibility of a recent peri- and parapatric speciation, in these two African species, including through an improved taxon sampling. *Nymphaea thermarum* has been proposed as a model species for the early evolution of traits in flowering plants (Povilus et al., 2020). The plastome of the individual of *N. cf. capensis* from Botswana differs from that of *N. thermarum*/*N. heudelotii* by only two nucleotide substitutions and approximately a dozen base pair insertions/deletions (primarily in simple sequence repeat regions and microsatellites), which indicates a very recent separation of this species from its West African congeners. Clearly, more taxonomic research on the African species of subgenus *Brachyceras* is needed.

Phylogenetic relationships within subgenus Nymphaea

Our phylogenetic reconstructions indicated some of the relationships within the autonymic subgenus of *Nymphaea*, which includes the type, *N. alba*. Specifically, we found that the North American species of the crown group of *N.* subg. *Nymphaea* (i.e. *N. mexicana* and *N. odorata*) were more closely related to each other than to the mainly Eurasian species *N. alba*. Previous molecular phylogenetic studies were inconsistent on these relationships and supported either *N. mexicana* (Borsch et al., 2007) or *N. odorata* (Volkova et al., 2010) as the earliest diverging lineage within the subgenus, although a split into a North American and a Eurasian-boreal subclade had been proposed by Borsch et al. (2014) based on nrITS sequence data. A more comprehensive taxon sampling is required in future investigations to clarify the relationships in *Nymphaea* subg. *Nymphaea*.

Importance of a motif-based approach in nucleotide sequence alignment

Accommodating sequence motifs during MSA is an important but algorithmically challenging task and has not yet been adequately accomplished in the context of molecular phylogenetic studies (Dijkstra et al., 2018).

While different methods for the algorithmic detection of conserved sequence motifs in genomic sequences have been developed, including the use of position-specific score matrices or hidden Markov models (e.g. D'haeseleer, 2006; Grant et al., 2011), most merely aim to identify orthologous sequence motifs in a sequence alignment, oftentimes through the inference of custom substitution matrices (Hashim et al., 2019). Some of these *de novo* motif discovery algorithms can take the phylogenetic relationships among the input sequences into account but nonetheless rely on the use of precalculated sets of common sequence motifs (e.g. Arnold et al., 2012). By contrast, only a few MSA algorithms have been developed to attempt the opposite process of generating improved sequence alignments if given a set of conserved sequence motifs—a process that is commonly referred to as “motif-aware” sequence alignment (Lelieveld et al., 2016). Most motif-aware alignment algorithms require an *a priori* knowledge of the approximate nucleotide sequence of the motifs expected among the input sequences (e.g. Dijkstra et al., 2018). However, this prerequisite renders the application of such algorithms in phylogenetic investigations impracticable, as the nucleotide sequences of phylogenetic investigations are typically not characterized individually, especially when mass-produced via high-throughput sequencing (e.g. Dylus et al., 2023). Owing to the lack of suitable alignment strategies, the application of motif-aware algorithms has, consequently, been very rare in phylogenetic investigations.

Relevance of motif-based sequence alignment to plastid phylogenomics

Alignment strategies that accommodate length-variable sequence motifs during MSA are a high priority for plastid phylogenomics investigations but are currently limited to the adjustment of software-generated alignments. The average plastid genome exhibits a high density of small sequence motifs, particularly among its noncoding regions, but despite their complex mutational dynamics, many of these regions are commonly used in plant phylogenetic investigations (Morton, 2003; Borsch and Quandt, 2009). The high proportion of small inversions as well as insertions and deletions among plastid sequence motifs (e.g. Orton et al., 2017) exacerbates the challenge of aligning their sequences across species. While these microstructural mutations exhibit recurring mutational patterns owing to their structural or functional constraints (Kelchner, 2000), they do not typically display consistent substitution rates and, thus, cannot be modelled via default position-specific score matrices or hidden character states, as employed in many motif-aware alignment algorithms (Dijkstra et al., 2018). In fact, no algorithm or automated MSA strategy known to us

has so far achieved a motif-aware alignment of the noncoding regions of the plastid genome. While MSA algorithms that employ machine learning and have been trained on sequence motif-rich datasets may be able to align plastid sequence motifs that are prone to microstructural mutations in the future (e.g. Petti et al., 2023), contemporary investigations are limited to a targeted adjustment of the output alignments (Löhne and Borsch, 2005).

In the absence of suitable software tools, plastid phylogenomic studies should conduct manual inspections and motif-based adjustments of software-generated sequence alignments. With the help of such inspections, microstructural mutations can be recognized; subsequent alignment adjustments will then improve their phylogenetic encoding in the analysis matrix. Small sequence inversions, for example, should be manually re-inverted, re-aligned with the other sequences and coded as a single-step event in a supplementary indel matrix that is included during phylogenetic reconstruction (Simmons and Ochoterena, 2000). Empirical examples of this alignment inspection and adjustment process are illustrated in Fig. 1. Microstructural mutations that are left unchanged would have a high risk of biasing the inference of nucleotide substitution rates and, by extension, the eventual phylogenetic tree. Despite the large amounts of sequence data involved, the present investigation placed a strong emphasis on the motif-based adjustment of software-generated sequence alignments to improve the phylogenetic encoding of these difficult-to-align regions.

Impact of alignment adjustments on phylogenetic inference

Our assessment of the impact of motif-based alignment adjustments on the validity of the alignments as well as the subsequent phylogenetic inferences indicated a general reduction in alignment length and variability but also in their level of homoplasy. These findings are in line with the expectation that a visual inspection of a sequence alignment for optimal positional homology, followed by a motif-based adjustment, reduces both the length and the overall variability of an alignment but also its level of homoplasy (Morrison et al., 2015). Unless the reduction of the number of PIS negatively affects the resolution of the phylogenetic inferences, such alignment adjustments are likely to lead to more reliable phylogenetic conclusions than studies without them (Simmons et al., 2010). Indeed, our alignment adjustments consistently reduced the level of homoplasy in the matrix and occasionally reduced the number of PIS per alignment to a degree that the phylogenetic reconstructions were impacted. Moreover, the improvements in positional homology achieved through our alignment adjustments were likely to have

been amplified by a better fit of the nucleotide substitution model to the actual sequence data during phylogenetic tree inference (e.g. Du et al., 2019).

The exact impact of motif-based alignment adjustments on phylogenetic tree inference is illustrated in our empirical examples in Fig. 1. The first intron of *clpP*, for example, exhibits a small SSR in the outgroup taxon (i.e. *Cabomba caroliniana*), yet the software-driven MSA did not recognize this sequence motif as a repeat (Fig. 1b). While a phylogenetic tree reconstruction for this intron without a manual inspection and adjustment of the software-generated alignment would not be likely to have resulted in a different tree topology, it would most probably have exaggerated the number of autapomorphic characters attributed to the outgroup and, thus, artificially increased the branch length between ingroup and outgroup. Likewise, in the intergenic spacer between the genes *atpH* and *atpI*, the software-driven MSA inferred a partially overlapping set of insertions and/or deletions, where likely only two small, nonoverlapping indels of 5 bp length each exist (Fig. 1c). Depending on the applied algorithm for indel coding, a phylogenetic reconstruction without a prior alignment adjustment would be likely to have resulted in the inference of different phylogenetic relationships than in the presence of such adjustments.

Our results are in line with those of Escobari et al. (2021), who found that motif-based adjustments of locus-wise alignments of complete plastid genome sequences were instrumental in the recovery of phylogenetic relationships in a group of closely related South American sunflowers. Their alignment adjustments caused a general reduction in the homoplasy level of their sequence matrices, which led to changes in their phylogenetic reconstructions; similar to the results in our investigation, some of the differences in tree topology attributable to alignment adjustment were even found to be statistically significant (Escobari et al., 2021). The application of manual alignment adjustments has also been practiced in several other plastid phylogenomic studies. Leebens-Mack et al. (2005), for example, manually adjusted software-generated sequence alignments in an initial plastid phylogenomic study on early-diverging flowering plants to account for suboptimal homology statements by the alignment algorithm. Specifically, they found that, unless adjusted, indel mutations affected the reconstruction of the phylogenetic position of *Amborella* and the water-lilies relative to other angiosperms and, thus, generated a binary indel matrix for additional parsimony-based reconstructions. A similar approach was taken by Ma et al. (2014), who manually adjusted sequence alignments of complete plastomes generated by MAFFT and then removed all sequence inversions of a size between 3 and 26 bp that were identified through visual alignment inspection;

their rationale behind that removal was that small inversions were prone to homoplasy and could mislead phylogenetic inference.

The topological differences among the inferred phylogenetic trees of this investigation that are attributable to alignment adjustments are small but nonetheless statistically significant. Specifically, the reconstructions conducted before alignment inspection and adjustment identified a sister relationship between *Nymphaea* cf. *capensis* and *Nymphaea* × *daubenyana* (Fig. 5a), as did all reconstructions in which indels were coded (Fig. 6a,b); the same phylogenetic reconstructions conducted after motif-based alignment adjustment, however, did not recover this relationship and instead suggested that *Nymphaea* × *daubenyana* was the earliest divergent taxon within the clade of West African species (Fig. 5b). These different phylogenetic positions may additionally be connected to the hybrid nature of *Nymphaea* × *daubenyana*, which constitutes an interspecific hybrid between *N. micrantha*, as evidenced by the leaf vivipary it shares only with that species, and another unknown species of subgenus *Brachyceras* (Heine and Mabberley, 1986).

Our evaluation of the impact of alignment adjustment on the inference of best-fitting nucleotide substitution models produced complex and largely idiosyncratic results (Tables S2–S4), preventing us from drawing clear conclusions on any underlying patterns of cause and effect. First, the differences in log-likelihoods of model fit before and after alignment adjustment did not coincide with most of the cases of divergent nucleotide substitution models: despite large differences in the log-likelihood values of the AIC, best-fitting nucleotide substitution models were often identical before and after alignment adjustment (e.g. gene *ycf1*), whereas even minor differences in the log-likelihood values occasionally coincided with divergent best-fitting substitution models (e.g. the intergenic spacer between genes *trnY-GTA* and *trnE-TTC*). Alignment length appeared to be a factor relevant to log-likelihood value differences before and after alignment adjustment, but no clear pattern could be identified. Second, significant differences in tree topology before and after alignment adjustment did not coincide with most of the cases of divergent nucleotide substitution models: none of the coding regions, only 13 of the intergenic spacers (i.e. 12% of total), and only two of the introns (i.e. 11% of total) that exhibited significant topology differences in their best ML trees also showed different best-fitting substitution models. The argument that significantly different tree topologies before and after alignment adjustment are associated with, or even caused by, changes in the best-fitting nucleotide substitution models is, thus, not supported by our findings. Instead, it seems that our motif-based alignment adjustments have a complex pattern of

impact on the subsequent phylogenetic inferences that is only partially reflected by changes in the nucleotide substitution models. Following these and the observations of Abadi et al. (2019), we employed the most parameter-rich nucleotide substitution model (GTR + G + I) for our phylogenetic reconstructions.

Ancestral character states of plastid sequence inversions

Our ancestral character state reconstructions of the 12 sequence inversions indicated that most of these inversions exhibit idiosyncratic patterns of sequence evolution (Fig. 7). A closer examination of the nucleotide sequences that flank these inversions revealed the presence of conserved palindromic sequence motifs in many of them, a phenomenon that causes these inversions to form stem-loop hairpin structures and which stabilizes their mRNA product upon transcription (Kim and Lee, 2005). The inversion in the *psbZ-trnG-GCC* intergenic spacer (inversion 7), for example, is located between a poly-C and a complementary poly-G microsatellite and comprises a total of 57 nucleotides. Its directionality cannot be inferred from our analyses, as the ancestral state reconstruction inferred only a single character state transition: between the ingroup and the outgroup. We also found that the inversions located in the terminal parts of longer and, thus, more stable hairpin-structures such as those found in the intergenic spacers *psbT-psbN* (inversion 6) and *trnH-psbA* (inversion 10) are considerably more homoplastic than other inversions, which is in line with the general model of hairpin-mediated mutational mechanisms (Kelchner and Wendel, 1996) as well as empirical evidence from angiosperms (Graham et al., 2000) and mosses (Hernandez-Maqueda et al., 2008). A high level of homoplasy in hairpin-mediated inversions in the *psbA-trnH* spacer was also reported from other plant genera (e.g. Degtjareva et al., 2012), which limits its use as a DNA barcode. The three species of *Victoria*, however, are characterized by two synapomorphic inversions (i.e. inversions 9 and 11) that have occurred without the formation of prominent palindromic sequence structures. Overall, small sequence inversions of the plastid genome do not typically exhibit a large contribution to the dominant evolutionary signal of their genomes but rather represent idiosyncratic and often homoplastic signal that can hamper MSA and character interpretation during phylogeny reconstruction (Kim and Lee, 2005). The identification of such sequence inversions during a visual inspection of software-generated sequence alignments, therefore, represents an important step in plastid phylogenomic analyses.

Conclusion

The results of this investigation corroborated the evolutionary relationships of Nymphaeaceae reported

by previous studies and clarified several relationships that were so far uncertain. Specifically, our analyses resolved several internal relationships of the Neotropical clade of subgenus *Hydrocallis* for the first time and increased the confidence into multiple previously reported relationships for other subgenera of *Nymphaea*. Hence, considerable molecular phylogenetic evidence now exists to conclude that (i) each of the five subgenera of *Nymphaea* is monophyletic, (ii) subgenus *Nymphaea* is sister to the rest of the genus, and (iii) the genera *Euryale* and *Victoria* are sister to a clade formed by the subgenera *Hydrocallis* and *Lotos*, rendering *Nymphaea* paraphyletic in its current circumscription. Moreover, the results of this investigation highlighted the importance of motif-based alignment inspections and adjustments in the analysis of plastid phylogenomic sequence data. The finding that such alignment adjustments improve the positional homology among plastid sequence alignments, and, by extension, the subsequent phylogenetic reconstructions, is in line with multiple similar studies. However, the observation that the results of some of our phylogenetic reconstructions were significantly different after motif-based alignment adjustment than before is new and highlights the issue of alignment accuracy in plastid phylogenomic analyses. A detailed inspection and, where necessary, motif-based adjustment of software-generated DNA sequence alignments should, thus, be considered a common standard rather than a nuisance in plastid phylogenomic studies.

Acknowledgements

The authors thank Sabine Scheel and Cathrin Schierenbeck of the Freie Universität Berlin as well as Gabriele Dröge of the Botanischer Garten und Botanisches Museum Berlin (BGBM) for assistance with DNA isolation and the cataloguing and submission of DNA isolates to the DNA bank of the BGBM. The authors acknowledge the high-performance computing services of the ZEDAT of the Freie Universität Berlin and of the Beocat Research Cluster at Kansas State University for providing allocations of computing time. The authors further acknowledge that the software TNT was made available with the sponsorship of the Willi Hennig Society. This manuscript constitutes part of a thesis by JAR toward a Master of Science degree at the Freie Universität Berlin. Open Access funding enabled and organized by Projekt DEAL.

Conflict of interest

The authors declare that they have no competing interests.

Data availability statement

The data that support the findings of this study are openly available in Zenodo at <https://doi.org/10.5281/zenodo.7860937>, reference number 7860937.

References

- Abadi, S., Azouri, D., Pupko, T. and Mayrose, I., 2019. Model selection may not be a mandatory step for phylogeny reconstruction. *Nat. Commun.* 10, 934.
- Akaike, H., 1974. A new look at the statistical model identification. *IEEE Trans Automat Contr* 19, 716–723.
- Arnold, P., Erb, I., Pachkov, M., Molina, N. and van Nimwegen, E., 2012. MotEvo: Integrated bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences. *Bioinformatics* 28, 487–494.
- Baldwin, E., McNair, M. and Leebens-Mack, J., 2023. Rampant chloroplast capture in *Sarracenia* revealed by plastome phylogeny. *Front. Plant Sci.* 14, 1237749.
- Barbosa, T., Trad, R., Bajay, M., Zucchi, M. and do Carmo E. do Amaral, M., 2018. Reestablishment of *Cabomba schwartzii* (Cabombaceae), an aquatic plant species endemic to the Brazilian amazon. *Phytotaxa* 367, 245–255.
- Bolger, A., Lohse, M. and Usadel, B., 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120.
- Borsch, T. and Quandt, D., 2009. Mutational dynamics and phylogenetic utility of noncoding chloroplast DNA. *Plant Systematics and Evolution* 282, 169–199.
- Borsch, T., Hilu, K., Quandt, D., Wilde, V., Neinhuis, C. and Barthlott, W., 2003. Noncoding plastid *trnT-trnF* sequences reveal a well resolved phylogeny of basal angiosperms. *J. Evol. Biol.* 16, 558–576.
- Borsch, T., Hilu, K., Wiersema, J., Löhne, C., Barthlott, W. and Wilde, V., 2007. Phylogeny of *Nymphaea* (Nymphaeaceae): Evidence from substitutions and microstructural changes in the chloroplast *trnT-trnF* region. *International Journal of Plant Sciences* 168, 639–671.
- Borsch, T., Löhne, C. and Wiersema, J., 2008. Phylogeny and evolutionary patterns in Nymphaeales: Integrating genes, genomes and morphology. *Taxon* 57, 1052–1081.
- Borsch, T., Löhne, C., Mbaye, M. and Wiersema, J., 2011. Towards a complete species tree of *Nymphaea*: Shedding further light on subg. *Brachyceras* and its relationships to the Australian waterlilies. *Telopea* 13, 193–217.
- Borsch, T., Wiersema, J., Hellquist, C., Löhne, C. and 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.
- Chatzou, M., Magis, C., Chang, J.-M., Kemena, C., Bussotti, G., Erb, I. and Notredame, C., 2016. Multiple sequence alignment modeling: Methods and applications. *Brief. Bioinform.* 17, 1009–1023.
- Darriba, D., Posada, D., Kozlov, A., Stamatakis, A., Morel, B. and Flouri, T., 2019. ModelTest-NG: A new and scalable tool for the selection of DNA and protein evolutionary models. *Mol. Biol. Evol.* 37, 291–294.
- Dejtjareva, G., Logacheva, M., Samigullin, T. and Terentjeva, E., 2012. Organisation of chloroplast *psbA-trnH* spacer in dicotyledonous angiosperms of the family *Umbelliferae*. *Biochemistry (Mosc.)* 77, 1056–1064.
- D'haeseleer, P., 2006. How does DNA sequence motif discovery work? *Nat. Biotechnol.* 24, 959–961.
- Dierckxsens, N., Mardulyn, P. and Smits, G., 2017. NOVOPlasty: De novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* 45, e18.

- Dijkstra, M., Bawono, P., Abeln, S., Feenstra, K., Fokkink, W. and Heringa, J., 2018. Motif-aware PRALINE: Improving the alignment of motif regions. *PLoS Comput. Biol.* 14, e1006547.
- Du, Y., Wu, S., Edwards, S. and Liu, L., 2019. The effect of alignment uncertainty, substitution models and priors in building and dating the mammal tree of life. *BMC Evol. Biol.* 191, 203.
- Dylus, D., Altenhoff, A., Majidian, S., Sedlazeck, F. and Dessimoz, C., 2023. Inference of phylogenetic trees directly from raw sequencing reads using Read2Tree. *Nat. Biotechnol.* 42, 139–147.
- Escobari, B., Borsch, T., Quedensley, T. and Gruenstaeudl, M., 2021. Plastid phylogenomics of the Gynoxoid group (Senecioneae, Asteraceae) highlights the importance of motif-based sequence alignment amid low genetic distances. *Am. J. Bot.* 108, 2235–2256.
- Farris, J., 1989. The retention index and the rescaled consistency index. *Cladistics* 5, 417–419.
- Fischer, E., 1988. Beiträge zur Flora Zentralafrikas, I. Eine neue *Nymphaea* sowie ein neuer *Streptocarpus* aus Rwanda. *Feddes Repertorium* 99, 385–390.
- Giorgashvili, E., Reichel, K., Caswara, C., Kerimov, V., Borsch, T. and Gruenstaeudl, M., 2022. Software choice and sequencing coverage can impact plastid genome assembly – A case study in the narrow endemic *Calligonum bakuense*. *Front. Plant Sci.* 13, 779830.
- Glanz, S. and Kueck, U., 2009. Trans-splicing of organelle introns – A detour to continuous RNAs. *Bioessays* 31, 921–934.
- Goloboff, P. and Morales, M., 2023. TNT version 1.6, with a graphical interface for MacOS and Linux, including new routines in parallel. *Cladistics* 39, 144–153.
- Graham, S., Reeves, P., Burns, A. and Olmstead, R., 2000. Microstructural changes in noncoding chloroplast DNA: Interpretation, evolution, and utility of indels and inversions in basal angiosperm phylogenetic inference. *International Journal of Plant Sciences* 161, S83–S96.
- Grant, C., Bailey, T. and Noble, W., 2011. FIMO: Scanning for occurrences of a given motif. *Bioinformatics* 27, 1017–1018.
- Greiner, S., Lehwark, P. and Bock, R., 2019. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: Expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* 47, W59–W64.
- Gruenstaeudl, M., 2019. Why the monophyly of Nymphaeaceae currently remains indeterminate: An assessment based on genome-wide plastid phylogenomics. *Plant Syst. Evol.* 305, 827–836.
- Gruenstaeudl, M., Nauheimer, L. and Borsch, T., 2017. Plastid genome structure and phylogenomics of Nymphaeales: Conserved gene order and new insights into relationships. *Plant Syst. Evol.* 303, 1251–1270.
- Gruenstaeudl, M., Gerschler, N. and Borsch, T., 2018. Bioinformatic workflows for generating complete plastid genome sequences - an example from *Cabomba* (Cabombaceae) in the context of the phylogenomic analysis of the water-lily clade. *Life* 8, 25.
- Hashim, F., Mabrouk, M. and Al-Atabany, W., 2019. Review of different sequence motif finding algorithms. *Avicenna J. Med. Biotechnol.* 11, 130–148.
- He, D., Gichira, A., Li, Z., Nzei, J., Guo, Y., Wang, Q. and Chen, J., 2018. Intergeneric relationships within the early-diverging angiosperm family Nymphaeaceae based on chloroplast phylogenomics. *Int. J. Mol. Sci.* 19, 3780.
- Heine, H. and Maberley, D., 1986. An Oxford waterlily. *The Kew Magazine* 3, 167–175.
- Hernandez-Maqueda, R., Quandt, D., Werner, O. and Munuoz, J., 2008. Phylogenetic relationships and generic classification of the Grimmiaceae. *Mol. Phylogenet. Evol.* 46, 863–877.
- Hickson, R., Simon, C., Cooper, A., Spicer, G., Sullivan, J. and Penny, D., 1996. Conserved sequence motifs, alignment, and secondary structure for the third domain of animal 12S rRNA. *Mol. Biol. Evol.* 13, 150–169.
- Hickson, R., Simon, C. and Perrey, S., 2000. The performance of several multiple-sequence alignment programs in relation to secondary-structure features for an rRNA sequence. *Mol. Biol. Evol.* 17, 530–539.
- Hilu, K., Borsch, T., Müller, K., Soltis, D., Soltis, P., Savolainen, V., Chase, M., Powell, M., Alice, L., Evans, R., Sauquet, H., Neinhuis, C., Slotta, T., Rohwer, J., Campbell, C. and Chatrou, L., 2003. Angiosperm phylogeny based on *matK* sequence information. *Am. J. Bot.* 90, 1758–1776.
- Jin, J., Yu, W., Yang, J., Song, Y., dePamphilis, C., Yi, T. and Li, D., 2020. GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* 21, 241.
- Katoh, K. and Standley, D., 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780.
- Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P. and Drummond, A., 2012. Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649.
- Kelchner, S., 2000. The evolution of non-coding chloroplast DNA and its application in plant systematics. *Ann. Mo. Bot. Gard.* 87, 482–498.
- Kelchner, S. and Wendel, J., 1996. Hairpins create minute inversions in non-coding regions of chloroplast DNA. *Curr. Genet.* 30, 259–262.
- Kim, K. and Lee, H., 2005. Widespread occurrence of small inversions in the chloroplast genomes of land plants. *Mol. Cells* 19, 104–113.
- Kluge, A. and Farris, J., 1969. Quantitative phyletics and the evolution of anurans. *Syst. Zool.* 18, 1–32.
- Korotkova, N., Nauheimer, L., Ter-Voskanyan, H., Allgaier, M. and Borsch, T., 2014. Variability among the most rapidly evolving plastid genomic regions is lineage-specific: Implications of pairwise genome comparisons in *Pyrus* (Rosaceae) and other angiosperms for marker choice. *PLoS One* 9, e112998.
- Leebens-Mack, J., Raubeson, L., Cui, L., Kuehl, J., Fourcade, M., Chumley, T., Boore, J., Jansen, R. and dePamphilis, C., 2005. Identifying the basal angiosperm node in chloroplast genome phylogenies: Sampling one's way out of the Felsenstein zone. *Mol. Biol. Evol.* 22, 1948–1963.
- Lelieveld, S., Schütte, J., Dijkstra, M., Bawono, P., Kinston, S., Göttgens, B., Heringa, J. and Bonzanni, N., 2016. ConBind: Motif-aware cross-species alignment for the identification of functional transcription factor binding sites. *Nucleic Acids Res.* 44, e72.
- Les, D., Schneider, E., Padgett, D., Soltis, P., Soltis, D. and Zanis, M., 1999. Phylogeny, classification and floral evolution of water lilies (Nymphaeaceae; Nymphaeales): A synthesis of non-molecular, *rbcL*, *matK*, and 18S rDNA data. *Syst. Bot.* 24, 28–46.
- Lewis, P., 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* 50, 913–925.
- Liang, P., Saqib, H., Zhang, X., Zhang, L. and Tang, H., 2018. Single-base resolution map of evolutionary constraints and annotation of conserved elements across major grass genomes. *Genome Biol. Evol.* 10, 473–488.
- de Lima, C. and Guilietti, A., 2013. *Nymphaea vanildae* (Nymphaeaceae): A new species from the Caatinga in Brazil. *Phytotaxa* 134, 42–48.
- de Lima, C., Machado, I. and Guilietti, A., 2021. Nymphaeaceae of Brasil. *Sitientibus Sér. Ciên. Biol.* 21, 21–71.
- Liu, L.-X., Du, Y.-X., Folk, R., Wang, S.-Y., Soltis, D., Shang, F.-D. and Li, P., 2020. Plastome evolution in Saxifragaceae and multiple plastid capture events involving *Heuchera* and *Tiarella*. *Front. Plant Sci.* 11, 361.
- Löhne, C. and Borsch, T., 2005. Molecular evolution and phylogenetic utility of the *petD* group II intron: A case study in basal angiosperms. *Mol. Biol. Evol.* 22, 317–332.
- Löhne, C., Borsch, T. and Wiersema, J., 2007. Phylogenetic analysis of Nymphaeales using fast-evolving and noncoding chloroplast markers. *Bot. J. Linn. Soc.* 154, 141–163.

- Löhne, C., Borsch, T., Jacobs, S., Hellquist, C. and Wiersema, J., 2008a. Nuclear and plastid DNA sequences reveal complex reticulate patterns in Australian water-lilies (*Nymphaea* subgenus *Anephyta*, Nymphaeaceae). *Aust. Syst. Bot.* 21, 229–250.
- Löhne, C., Yoo, M., Borsch, T., Wiersema, J., Wilde, V., Bell, C., Barthlott, W., Soltis, D. and Soltis, P., 2008b. Biogeography of Nymphaeales: Extant patterns and historical events. *Taxon* 57, 1123–1146.
- Löhne, C., Wiersema, J. and Borsch, T., 2009. The unusual *Ondinea*, actually just another Australian water-lily of *Nymphaea* subgenus *Anephyta* (Nymphaeaceae). *Willdenowia* 39, 55–58.
- Long, H., Li, M. and Fu, H., 2016. Determination of optimal parameters of MAFFT program based on BALiBASE3.0 database. *SpringerPlus* 5, 736.
- Ma, P.-F., Zhang, Y.-X., Zeng, C.-X., Guo, Z.-H. and Li, D.-Z., 2014. Chloroplast phylogenomic analyses resolve deep-level relationships of an intractable bamboo tribe Arundinarieae (Poaceae). *Syst. Biol.* 63, 933–950.
- Morrison, D., 2006. Multiple sequence alignment for phylogenetic purposes. *Aust. Syst. Bot.* 19, 479–539.
- Morrison, D., 2008. A framework for phylogenetic sequence alignment. *Plant Syst. Evol.* 282, 127–149.
- Morrison, D., 2009. Why would phylogeneticists ignore computerized sequence alignment? *Syst. Biol.* 58, 150–158.
- Morrison, D., Morgan, M. and Kelchner, S., 2015. Molecular homology and multiple-sequence alignment: An analysis of concepts and practice. *Aust. Syst. Bot.* 28, 46–62.
- Morton, B., 2003. The role of context-dependent mutations in generating compositional and codon usage bias in grass chloroplast DNA. *J. Mol. Evol.* 56, 616–629.
- Moseley, M., 1961. Morphological studies of the Nymphaeaceae II. The flower of *Nymphaea*. *Bot. Gazette* 122, 233–259.
- Moseley, M., Schneider, E. and Williamson, P., 1993. Phylogenetic interpretations from selected floral vasculature characters in the Nymphaeaceae sensu lato. *Aquatic Botany* 44, 325–342.
- Mower, J. and Vickrey, T., 2018. Structural diversity among plastid genomes of land plants. *Adv. Bot. Res.* 85, 263–292.
- Müller, J., Müller, K., Neinhuis, C. and Quandt, D., 2010. PhyDE: Phylogenetic Data Editor. Available from: <http://www.phyde.de/>. Accessed 19-Aug-2020.
- Ogden, T. and Rosenberg, M., 2006. Multiple sequence alignment accuracy and phylogenetic inference. *Syst. Biol.* 55, 314–328.
- Orton, L., Burke, S., Wysocki, W. and Duvall, M., 2017. Plastid phylogenomic study of species within the genus *Zea*: Rates and patterns of three classes of microstructural changes. *Curr. Genet.* 63, 311–323.
- Paradis, E. and Schliep, K., 2018. Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528.
- Pereira, W., Knaack, S., Chakraborty, S., Conde, D., Folk, R., Triozzi, P., Balmant, K., Dervinis, C., Schmidt, H., Ané, J.-M., Roy, S. and Kirst, M., 2022. Functional and comparative genomics reveals conserved noncoding sequences in the nitrogen-fixing clade. *New Phytol.* 234, 634–649.
- Petti, S., Bhattacharya, N., Rao, R., Dauparas, J., Thomas, N., Zhou, J., Rush, A., Koo, P. and Ovchinnikov, S., 2023. End-to-end learning of multiple sequence alignments with differentiable Smith-Waterman. *Bioinformatics* 39, btac724.
- Phillips, A., Janies, D. and Wheeler, W., 2000. Multiple sequence alignment in phylogenetic analysis. *Mol. Phylogenet. Evol.* 16, 317–330.
- Polhill, R. and Verdcourt, B., 1989. *Flora of Tropical East Africa: Nymphaeaceae*. A.A.Balkema, Rotterdam.
- Povilus, R., DaCosta, J., Satyaki, P., Moeglein, M., Jaenisch, J., Xi, Z., Mathews, S., Gehring, M., Davis, C. and Friedman, W., 2020. Water lily (*Nymphaea thermarum*) genome reveals variable genomic signatures of ancient cambium losses. *Proc. Natl. Acad. Sci. U.S.A.* 117, 8649–8656.
- R Development Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. Available from: <http://www.r-project.org/>.
- Rambaut, A., Drummond, A., Xie, D., Baele, G. and Suchard, M., 2018. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* 67, 901–904.
- Ronquist, F. and Huelsenbeck, J., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572–1574.
- Salinas, N. and Little, D., 2014. 2matrix: A utility for indel coding and phylogenetic matrix concatenation. *Appl. Plant Sci.* 2, apps.1300083.
- Schliep, K., 2011. Phangorn: Phylogenetic analysis in R. *Bioinformatics* 27, 592–593.
- Schneider, E., Tucker, S. and Williamson, P., 2003. Floral development in the Nymphaeales. *Int. J. Plant Sci.* 164, S279–S292.
- Sela, I., Ashkenazy, H., Katoh, K. and Pupko, T., 2015. GUIDANCE2: Accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res.* 43, W7–W14.
- Shaw, J., Lickey, E., Schilling, E. and Small, R., 2007. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: The tortoise and the hare III. *Am. J. Bot.* 94, 275–288.
- Shimodaira, H., 2002. An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* 51, 492–508.
- Shimodaira, H. and Hasegawa, M., 2001. CONSEL: For assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17, 1246–1247.
- Simmons, M. and Ochoterena, H., 2000. Gaps as characters in sequence-based phylogenetic analyses. *Syst. Biol.* 49, 369–381.
- Simmons, M., Müller, K. and Norton, A., 2010. Alignment of, and phylogenetic inference from, random sequences: The susceptibility of alternative alignment methods to creating artifactual resolution and support. *Mol. Phylogenet. Evol.* 57, 1004–1016.
- Smith, S., Walker-Hale, N. and Walker, J., 2020. Intra-genomic conflict in phylogenomic data sets. *Mol. Biol. Evol.* 37, 3380–3388.
- Smith, L., Magdalena, C., Przelomska, N., Perez-Escobar, O., Melgar-Gomez, D., Beck, S., Negro, R., Mian, S., Leitch, I., Dodsworth, S., Maurin, O., Ribero-Guardia, G., Salazar, C., Gutierrez-Sibauty, G., Antonelli, A. and Monro, A., 2022. Revised species delimitation in the giant water lily genus *Victoria* (Nymphaeaceae) confirms a new species and has implications for its conservation. *Front. Plant Sci.* 13, 883151.
- Stamatakis, A., 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313.
- Stamatakis, A., Hoover, P. and Rougemont, J., 2008. A rapid bootstrap algorithm for the RAxML web servers. *Syst. Biol.* 57, 758–771.
- Tan, G., Muffato, M., Ledergerber, C., Herrero, J., Goldman, N., Gil, M. and Dessimoz, C., 2015. Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. *Syst. Biol.* 64, 778–791.
- Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E., Fischer, A., Bock, R. and Greiner, S., 2017. GeSeq – Versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* 45, W6–W11.
- Volkova, P., Travnicek, P. and Brochmann, C., 2010. Evolutionary dynamics across discontinuous freshwater systems: Rapid expansions and repeated allopolyploid origins in the Palearctic white water-lilies (*Nymphaea*). *Taxon* 59, 483–494.
- Wiersema, J., 1987. A monograph of *Nymphaea* subgenus *Hydrocallis* (Nymphaeaceae). *Syst. Bot. Monogr.* 16, 1–112.
- Wong, K., Suchard, M. and Huelsenbeck, J., 2008. Alignment uncertainty and genomic analysis. *Science* 319, 473–476.
- Wu, M., Chatterji, S. and Eisen, J., 2012. Accounting for alignment uncertainty in phylogenomics. *PLoS One* 7, 1–10.
- Yang, Z., Kumar, S. and Nei, M., 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141, 1641–1650.

Zhang, L., Chen, F., Zhang, X., Li, Z., Zhao, Y., Lohaus, R., Chang, X., Dong, W., Ho, S., Liu, X., Song, A., Chen, J., Guo, W., Wang, Z., Zhuang, Y., Wang, H., Chen, X., Hu, J., Liu, Y., Qin, Y., Wang, K., Dong, S., Liu, Y., Zhang, S., Yu, X., Wu, Q., Wang, L., Yan, X., Jiao, Y., Kong, H., Zhou, X., Yu, C., Chen, Y., Li, F., Wang, J., Chen, W., Chen, X., Jia, Q., Zhang, C., Jiang, Y., Zhang, W., Liu, G., Fu, J., Chen, F., Ma, H., van de Peer, Y. and Tang, H., 2020. The water lily genome and the early evolution of flowering plants. *Nature* 577, 79–84.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1. Overview of alignment adjustments for all 203 coding and noncoding plastid genome regions employed for phylogenetic reconstruction.

Table S2. Overview of the best-fitting nucleotide substitution models and the log-likelihoods of model fit for each of the coding regions under study.

Table S3. Overview of the best-fitting nucleotide substitution models and the log-likelihoods of model fit for each of the intergenic spacers under study.

Table S4. Overview of the best-fitting nucleotide substitution models and the log-likelihoods of model fit for each of the introns under study.

Table S5. Overview of the number of parsimony-informative sites and significant differences in tree topology for all coding regions under study.

Table S6. Overview of the number of parsimony-informative sites and significant differences in tree topology for all intergenic spacers under study.

Table S7. Overview of the number of parsimony-informative sites and significant differences in tree topology for all introns under study.

Fig. S1. Phylogenetic relationships of Nymphaeaceae inferred under BI on the concatenated data set of all coding and noncoding plastome regions when indels are uncoded.

Fig. S2. Phylogenetic relationships of Nymphaeaceae inferred under BI on the concatenated data set of all coding and noncoding plastome regions when indels are coded.

Fig. S3. Phylogenetic relationships of Nymphaeaceae inferred under MP on the concatenated data set of all coding and noncoding plastome regions when indels are uncoded.

Fig. S4. Phylogenetic relationships of Nymphaeaceae inferred under MP on the concatenated data set of all coding and noncoding plastome regions when indels are coded.