 doi:10.20944/preprints201905.0002.v1

*Article*

# Why the monophyly of Nymphaeaceae currently remains indeterminate: An assessment based on gene-wise plastid phylogenomics

**Michael Gruenstaeudl** [1,*]

[1] Institut für Biologie, Freie Universität Berlin, 14195 Berlin, Germany; m.gruenstaeudl@fu-berlin.de
* Correspondence: m.gruenstaeudl@fu-berlin.de

Received: date; Accepted: date; Published: date

**Abstract:** The monophyly of Nymphaeaceae (water lilies) represents a critical question in understanding the evolutionary history of early-diverging angiosperms. A recent plastid phylogenomic investigation claimed new evidence for the monophyly of Nymphaeaceae, but its results could not be verified from the available data. Moreover, preliminary gene-wise analyses of the same dataset provided partial support for the paraphyly of the family. The present investigation aims to re-assess the previous conclusion of the monophyly of Nymphaeaceae under the same dataset and to determine the congruence of the phylogenetic signal across different plastome genes and data partition strategies. To that end, phylogenetic tree inference is conducted on each of 78 protein-coding plastome genes, both individually and upon concatenation, and under four data partitioning schemes. Moreover, the possible effects of various sequence variability and homoplasy metrics on the inference of specific phylogenetic relationships are tested using multiple logistic regression. Differences in the variability of polymorphic sites across codon positions are assessed using parametric and non-parametric analysis of variance. The results of the phylogenetic reconstructions indicate considerable incongruence among the different gene trees as well as the data partitioning schemes. The results of the multiple logistic regression tests indicate that the fraction of polymorphic sites of codon position 3 has a significant effect on the recovery of the monophyly of Nymphaeaceae. Taken together, these results indicate that the monophyly of Nymphaeaceae currently remains indeterminate, and that specific phylogenetic conclusions are strongly dependent on the precise plastome gene, data partitioning scheme, and codon position evaluated. In closing, I discuss the importance of archiving all data of an investigation in publicly accessible data repositories, along with sufficient details to replicate the published results, and provide recommendations on future plastid phylogenomic investigations of Nymphaeales.

**Keywords:** codon position; complete chloroplast; early-diverging angiosperms; monophyly; Nymphaeaceae; Nymphaeales; plastid genome; phylogenomics; repeatability

## 1. Introduction

The plant order Nymphaeales plays a pivotal role in our understanding of the evolutionary history of early-diverging angiosperms, with the phylogenetic position of the genus *Nuphar* being of particular interest. Upon the split between the earliest-diverging angiosperm *Amborella* and all other angiosperms, Nymphaeales likely represent the next branch-off within the flowering plants [1–3]. In its current taxonomic circumscription, Nymphaeales include three families with a total of eight genera: Nymphaeaceae (five genera; [4,5]), Cabombaceae (two genera; [6,7]), and Hydatellaceae (one genus; [8,9]). The species diversity of Nymphaeales is notably unequal: the paraphyletic genus *Nymphaea* comprises more than half of the species of the entire order ($>$ 50 sp.; [4,10]), whereas the genera *Brasenia* and *Cabomba* (both Cabombaceae) as well as *Victoria*, *Barclaya* and *Euryale* (all Nymphaeaceae) each comprise at most five species. The Hydatellaceae comprise a single genus with 12 species (*Trithuria*; [9]) and were identified as part of the water lily clade only recently [11,12]. The exact phylogenetic

 doi:10.20944/preprints201905.0002.v1



relationships within Nymphaeales remain controversial, primarily due to the uncertain position of the genus *Nuphar*. Two recent plastid phylogenomic investigations have recovered *Nuphar* apart from the other Nymphaeaceae [13,14], whereas older molecular phylogenetic studies have inferred the genus as the earliest diverging lineage of the family, albeit with mixed phylogenetic signal [4,15,16]. The precise phylogenetic position of *Nuphar* affects the taxonomic status of Nymphaeaceae as a monophyletic group (Fig. 1). A paraphyly of Nymphaeaceae would have considerable consequences for our understanding of macro- and micromorphological evolutionary processes within Nymphaeales and, by extension, the angiosperms as a whole [17–19]. Thus, the phylogenetic position of *Nuphar* and the question of monophyly of Nymphaeaceae is of considerable importance to evolutionary botany.

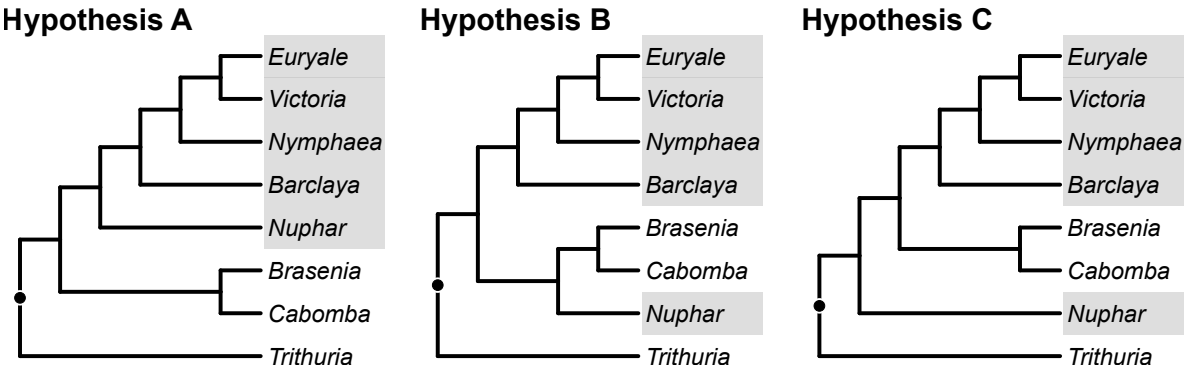


**Figure 1.** The three phylogenetic positions of *Nuphar* within Nymphaeales as recovered by previous investigations, displayed as separate phylogenetic hypotheses. The genera of Nymphaeaceae are highlighted in gray to visualize the placement of *Nuphar* under which Nymphaeaceae are monophyletic (i.e., hypothesis A) versus those placements under which the family is paraphyletic (i.e., hypotheses B and C).

A recent plastid phylogenomic investigation claimed new evidence for the monophyly of Nymphaeaceae but, upon closer examination, this evidence appears controversial and warrants a re-assessment. The recent study of He et al. [20] aimed to evaluate the intergeneric relationships within Nymphaeaceae using complete plastid genome sequences. It is the second investigation to employ complete plastid genome sequences for evaluating the evolutionary relationships within Nymphaeales and combined 12 previously published plastid genomes (as compiled in [13]) with five newly sequenced ones. Among the primary aims of [20] was the objective to identify "the ideal rooting group and using it to elucidate the phylogenetic position of *Nuphar* and delimit intergeneric relationships within Nymphaeaceae family" ([20], p.2). To do so, the authors extracted coding regions of the 17 plastid genomes under study and reconstructed the phylogenetic relationships to determine the position of *Nuphar* and, by extension, the monophyly of Nymphaeaceae. Simultaneously, the authors tested the inclusion of other previously published plastid genomes to identify a suitable outgroup. By using this strategy, [20] recovered the genus *Nuphar* as the earliest divergence within Nymphaeaceae when only members of Hydatellaceae were used as an outgroup and concluded that the family was, thus, monophyletic. In fact, the authors asserted that, based on their results, the phylogenetic position of *Nuphar* was now "ascertained" ([20], p.16). However, upon closer evaluation of the study of [20], several aspects of their phylogenetic analyses appear enigmatic. For example, essential information to examine or replicate their phylogenetic reconstructions is neither provided in their publication nor accessible through public databases. Moreover, preliminary results of gene-wise phylogenetic reconstructions on the same dataset recover Nymphaeaceae as paraphyletic under various genes, indicating that the conclusion of monophyly of Nymphaeaceae may be premature. Thus, a re-assessment of the phylogenetic conclusions reached by [20] is warranted.

The phylogenetic conclusions reached by [20] cannot be verified from the publicly available data. Several gaps concerning the phylogenetic analyses presented by [20] exist in the publicly accessible information. These gaps prevent other researchers from replicating their results and, by extension,

 doi:10.20944/preprints201905.0002.v1



verifying their phylogenetic conclusions. For example, the actual set of coding regions that [20] extracted from the 17 plastid genomes under study and employed for phylogeny reconstruction is currently indeterminate. As of 01-Mar-2019, this information is neither given in their publication nor has it been made available on publicly accessible databases such as TreeBASE [21] or Dryad (https://datadryad.org/; [22]). It was possible to obtain the relevant information in a direct email correspondence with the authors of [20], albeit with considerable confusion (see Methods). Similar uncertainty exists regarding the data partitioning schemes that were employed in several of their phylogenetic analyses. These schemes are not described beyond stating the software used for finding the optimal nucleotide substitution models of the partitions. Again, the publicly accessible information is insufficient for replicating the results and, by extension, verifying the conclusions. Moreover, potentially incorrect annotations of protein-coding sequences are present in some of the newly assembled plastid genomes of [20] as deposited to GenBank. These annotations contain internal stop codons that were not listed in their table of RNA editing sites and should, thus, have marked the end of their respective amino acid positions. If used without prior adjustment when extracting coding sequences for subsequent DNA alignment and phylogeny reconstruction, these annotations would have resulted in the inclusion of DNA sequences of incorrect length. However, since the DNA alignments analyzed by [20] have not been made publicly accessible either, it is not possible for other researchers to evaluate these concerns.

A re-assessment of the phylogenetic conclusions of [20] is also indicated because a preliminary, gene-wise analysis of the same dataset provided partial support for the paraphyly of Nymphaeaceae. In a preliminary evaluation of the results of [20], gene-wise reconstructions of the phylogenetic history of Nymphaeales were conducted using the same dataset and the same tree inference algorithms. Despite the methodological equality to the original analyses, these reconstructions did not unanimously support the phylogenetic relationships reported by [20]. Instead, they indicated that different plastome genes supported different phylogenetic relationships between the input taxa (Table 1). Among the ten protein-coding plastome genes with the highest level of parsimony-informative characters, six supported the placement of *Nuphar* as the earliest-diverging member of Nymphaeaceae and, by extension, the monophyly of the family (hypothesis A in Fig. 1). By contrast, three genes indicated that *Nuphar* was sister to Cabombaceae (hypothesis B in Fig. 1) and Nymphaeaceae, thus paraphyletic. One gene recovered *Nuphar* as the earliest diverging lineage of a clade comprising Cabombaceae and the remaining Nymphaeaceae, again rendering Nymphaeaceae paraphyletic (hypothesis C in Fig. 1). Given these results, the conclusion by [20] that the monophyly of Nymphaeaceae had been ascertained may be premature. Instead, the preliminary results suggest that a re-evaluation of the phylogenetic relationships of Nymphaeaceae and Cabombaceae with particular emphasis on gene-wise phylogenetic signal and fine-scale data partitioning is warranted.

In the present investigation, the same dataset of 17 complete plastid genomes as analyzed by [20] is re-evaluated to corroborate their conclusion of monophyly of Nymphaeaceae. Phylogenetic reconstructions on 78 protein-coding genes encoded in the plastid genome are conducted, both on a gene-by-gene basis as well as on a multi-gene alignment. This multi-gene alignment constitutes a concatenation of the 78 plastome genes and is analyzed under four different, a priori defined data partitioning schemes. In order to determine potential factors behind the conflicting phylogenetic signal between the different genes, a series of sequence variability and homoplasy metrics is calculated on each of the 78 genes, including the proportion of polymorphic sites for each codon position. Possible effects between any of these sequence variability/homoplasy metrics and the tree inference results for each gene are explored by using multiple logistic regression models. Moreover, differences in the variability of polymorphic sites across codon positions as well as the significance of these differences are assessed using parametric and non-parametric analysis of variance. Particular attention is given to the documentation of the individual analysis steps and their results. Specifically, each of the DNA alignments, partition schemes, and phylogenetic trees analyzed is made publicly available for future re-analysis.

 doi:10.20944/preprints201905.0002.v1



**Table 1.** Results of the preliminary evaluation of sequence variability and homoplasy metrics of ten protein-coding plastome genes as well as the congruence of their gene-wise phylogenetic reconstruction results with the phylogenetic hypotheses of Fig. 1. The genes evaluated represent the ten genes with the highest level of parsimony-informative characters among the total of 78 protein-coding plastome genes evaluated in this investigation. Abbreviations used: Polymorphic = fraction of polymorphic sites; Informative = fraction of parsimony-informative sites; CI = consistency index; RI = retention index; P-distance = maximum uncorrected p-distance; Hypo = hypothesis.

| Gene | Polymorphic | Informative | CI | RI | P-distance | HypoA | HypoB | HypoC |
|---|---|---|---|---|---|---|---|---|
| *ycf1* | 0.999 | 0.306 | 0.888 | 0.891 | 0.72 | 1 | 0 | 0 |
| *rps15* | 0.284 | 0.234 | 0.838 | 0.882 | 0.459 | 1 | 0 | 0 |
| *rpl32* | 0.386 | 0.216 | 0.94 | 0.932 | 0.497 | 1 | 0 | 0 |
| *atpE* | 0.241 | 0.214 | 0.879 | 0.897 | 0.436 | 0 | 1 | 0 |
| *ndhF* | 0.28 | 0.214 | 0.867 | 0.888 | 0.439 | 1 | 0 | 0 |
| *rps19* | 0.323 | 0.202 | 0.867 | 0.917 | 0.435 | 1 | 0 | 0 |
| *infA* | 0.28 | 0.199 | 0.934 | 0.946 | 0.456 | 0 | 0 | 1 |
| *matK* | 0.267 | 0.197 | 0.9 | 0.904 | 0.433 | 1 | 0 | 0 |
| *ccsA* | 0.262 | 0.192 | 0.898 | 0.912 | 0.424 | 0 | 1 | 0 |
| *rpoC2* | 0.292 | 0.188 | 0.911 | 0.922 | 0.421 | 0 | 1 | 0 |

The present investigation aims to identify possible factors behind the incongruent phylogenetic signal of different plastome genes of Nymphaeales using an expanded set of analyses compared to [20]. Through the application of gene-wise phylogenetic reconstructions and multivariate statistical analyses, this investigation does not merely replicate the analyses of [20]. Instead, the analyses performed here go beyond the original analyses and attempt to identify possible factors causing the incongruent phylogenetic signal between different plastome genes. Specifically, I aim to answer four different questions which enable the assessment of the phylogenetic conclusions of [20] and simultaneously point to future directions for reconstructing unresolved relationships within Nymphaeales. These questions are: (i) Are the results of phylogenetic inference performed on individual plastome genes congruent across genes? (ii) Do sequence variability and homoplasy metrics of the plastome genes have statistically significant effects on the results of phylogenetic inference performed on these genes? (iii) Are the results of phylogenetic inference performed on the multi-gene alignment congruent under different data partitioning schemes? (iv) Has the monophyly of Nymphaeaceae been ascertained given the plastid genome data analyzed by [20]?

## 2. Results

### *2.1. Incongruence among phylogenetic trees inferred under different genes*

The phylogenetic trees inferred under individual gene alignments were found to be incongruent across the different genes of the plastid genome (Table 2). Of the 78 genes analyzed, 30 supported phylogenetic hypothesis A, 23 supported hypothesis B, and 15 supported hypothesis C. Reconstructions on the remaining ten genes resulted either in phylogenetic trees that were incongruent with either of the three phylogenetic hypotheses or in trees that contained polytomies at the relevant nodes and, thus, did not allow the determination of congruence with either hypothesis. Hence, the majority of the 78 protein-coding plastome genes that were congruent with any of the three phylogenetic hypotheses postulated supported the paraphyly of Nymphaeaceae. The best ML trees for each of the 78 gene-wise reconstructions are displayed in FigureS1.

### *2.2. Effects between recovered phylogenetic relationships*

The relative model fit of different logistic regression models indicated that the fraction of polymorphic sites of codon position 3 had a significant effect on the recovery of the monophyly

 doi:10.20944/preprints201905.0002.v1



**Table 2.** Results of the gene-wise phylogenetic reconstructions under ML tree inference for each of the 78 protein-coding plastome genes under study. The genes are sorted in the same order as found in the actual plastid genomes, starting with the 5′ end of the large single copy region and using the plastid genome of *Victoria cruziana* (GenBank accession number KY001813) as reference. The row indicated by the name '78 CDS' presents the results of the phylogenetic reconstruction under the multi-gene alignment of the 78 protein-coding plastome genes. Abbreviations used: Hypo = hypothesis.

| # | Gene | HypoA | HypoB | HypoC | # | Gene | HypoA | HypoB | HypoC |
|---|---|---|---|---|---|---|---|---|---|
| 1 | *78 CDS* | 1 | 0 | 0 | 41 | *petG* | 0 | 0 | 0 |
| 2 | *psbA* | 0 | 0 | 0 | 42 | *psaJ* | 1 | 0 | 0 |
| 3 | *matK* | 1 | 0 | 0 | 43 | *rpl33* | 1 | 0 | 0 |
| 4 | *rps16* | 1 | 0 | 0 | 44 | *rps18* | 1 | 0 | 0 |
| 5 | *psbK* | 1 | 0 | 0 | 45 | *rpl20* | 0 | 1 | 0 |
| 6 | *psbI* | 1 | 0 | 0 | 46 | *psbB* | 1 | 0 | 0 |
| 7 | *atpA* | 0 | 1 | 0 | 47 | *psbT* | 1 | 0 | 0 |
| 8 | *atpF* | 0 | 1 | 0 | 48 | *psbN* | 0 | 1 | 0 |
| 9 | *atpH* | 1 | 0 | 0 | 49 | *psbH* | 0 | 1 | 0 |
| 10 | *atpI* | 0 | 1 | 0 | 50 | *petB* | 0 | 1 | 0 |
| 11 | *rps2* | 0 | 1 | 0 | 51 | *petD* | 0 | 1 | 0 |
| 12 | *rpoC2* | 0 | 1 | 0 | 52 | *rpoA* | 1 | 0 | 0 |
| 13 | *rpoC1* | 1 | 0 | 0 | 53 | *rps11* | 0 | 1 | 0 |
| 14 | *rpoB* | 0 | 1 | 0 | 54 | *rpl36* | 0 | 0 | 0 |
| 15 | *petN* | 0 | 0 | 0 | 55 | *infA* | 0 | 0 | 1 |
| 16 | *psbM* | 0 | 0 | 0 | 56 | *rps8* | 0 | 0 | 1 |
| 17 | *psbD* | 1 | 0 | 0 | 57 | *rpl14* | 1 | 0 | 0 |
| 18 | *psbC* | 0 | 1 | 0 | 58 | *rpl16* | 1 | 0 | 0 |
| 19 | *psbZ* | 0 | 1 | 0 | 59 | *rps3* | 0 | 0 | 1 |
| 20 | *rps14* | 0 | 0 | 1 | 60 | *rpl22* | 0 | 1 | 0 |
| 21 | *psaB* | 0 | 0 | 1 | 61 | *rps19* | 1 | 0 | 0 |
| 22 | *psaA* | 0 | 1 | 0 | 62 | *rpl2* | 0 | 0 | 1 |
| 23 | *ycf3* | 0 | 0 | 1 | 63 | *rpl23* | 0 | 0 | 0 |
| 24 | *rps4* | 1 | 0 | 0 | 64 | *ycf2* | 1 | 0 | 0 |
| 25 | *ndhJ* | 0 | 0 | 1 | 65 | *ndhB* | 1 | 0 | 0 |
| 26 | *ndhK* | 1 | 0 | 0 | 66 | *rps7* | 0 | 0 | 1 |
| 27 | *ndhC* | 0 | 0 | 1 | 67 | *rps12* | 0 | 1 | 0 |
| 28 | *atpE* | 0 | 1 | 0 | 68 | *ndhF* | 1 | 0 | 0 |
| 29 | *atpB* | 1 | 0 | 0 | 69 | *rpl32* | 1 | 0 | 0 |
| 30 | *rbcL* | 0 | 1 | 0 | 70 | *ccsA* | 0 | 1 | 0 |
| 31 | *accD* | 1 | 0 | 0 | 71 | *ndhD* | 0 | 0 | 1 |
| 32 | *psaI* | 0 | 1 | 0 | 72 | *psaC* | 1 | 0 | 0 |
| 33 | *ycf4* | 1 | 0 | 0 | 73 | *ndhE* | 1 | 0 | 0 |
| 34 | *cemA* | 0 | 0 | 1 | 74 | *ndhG* | 0 | 0 | 1 |
| 35 | *petA* | 0 | 0 | 1 | 75 | *ndhI* | 0 | 0 | 0 |
| 36 | *psbJ* | 0 | 1 | 0 | 76 | *ndhA* | 0 | 1 | 0 |
| 37 | *psbL* | 0 | 0 | 0 | 77 | *ndhH* | 1 | 0 | 0 |
| 38 | *psbF* | 0 | 0 | 0 | 78 | *rps15* | 1 | 0 | 0 |
| 39 | *psbE* | 0 | 0 | 1 | 79 | *ycf1* | 1 | 0 | 0 |
| 40 | *petL* | 0 | 0 | 0 | | | | | |

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 May 2019 doi:10.20944/preprints201905.0002.v1



of Nymphaeaceae during phylogeny inference (Table 3). Specifically, the addition of a parameter that specifies an effect of the fraction of polymorphic sites of codon position 3 on the recovery of phylogenetic hypothesis A was found to significantly increase the fit of the model (p-value $< 0.01$). Under the AIC, the difference between this model and the null model was considerably larger than for any other model comparison ($\Delta$AIC = 10.51). No other addition of effect parameters was found to increase model fit to the data significantly. The relative fit of the tested models under the AIC and the LRT, and the significance value of the likelihood ratio statistic of each model comparison in relation to a $\chi^2$ distribution are given in Table 3. The results of calculating the different sequence variability and homoplasy metrics as well as the fractions of polymorphic sites of the three codon positions across the 78 protein-coding plastome genes under study which are the basis for the tests of possible effects on the phylogenetic tree inference as well as the overall and pairwise comparisons of the mean fraction values are given in TableS1 and TableS2.

**Table 3.** Results of a series of multiple logistic regression tests in order to identify significant effects of different sequence variability/homoplasy metrics on the results of the gene-wise phylogenetic reconstructions. Displayed are the degrees of freedom, the difference between the AIC values, the LRT statistic, and the significance of the LRT statistic for each model comparison conducted. Asterisks indicate a significance of the LRT statistic at $p < 0.01$. For easier visualization, the display of the sequence variability and homoplasy metric added to the respective alternative model is sorted alphabetically. Abbreviations used: $\Delta$AIC = difference between the AIC values of the null and the alternative model; CI = consistency index; CodonOne = fraction of polymorphic sites of codon 1; CodonTwo = fraction of polymorphic sites of codon 2; CodonThree = fraction of polymorphic sites of codon 3; DF = degrees of freedom; GC = GC content; Hypo = hypothesis; Informative = fraction of parsimony-informative sites; Length = alignment length; LRT = likelihood ratio test statistic; n.a. = not applicable; P-distance = maximum uncorrected p-distance; Polymorphic = fraction of polymorphic sites; RI = retention index.

| | **HypoA** | | | | **HypoB** | | | | **HypoC** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Df | $\Delta$AIC | LRT | P>$\chi^2$ | Df | $\Delta$AIC | LRT | P>$\chi^2$ | Df | $\Delta$AIC | LRT | P>$\chi^2$ |
| *Null model* | 0 | 95.43 | n.a. | n.a. | 0 | 95.24 | n.a. | n.a. | 0 | 78.37 | n.a. | n.a. |
| CI | 1 | 96.36 | 1.07 | 0.30 | 1 | 95.59 | 1.65 | 0.20 | 1 | 80.04 | 0.33 | 0.57 |
| CodonOne | 1 | 96.31 | 1.11 | 0.29 | 1 | 97.24 | 0.00 | 0.98 | 1 | 80.27 | 0.10 | 0.75 |
| CodonThree | 1 | 105.94 | 12.51 | <0.01* | 1 | 97.23 | 0.01 | 0.95 | 1 | 79.56 | 0.81 | 0.37 |
| CodonTwo | 1 | 97.26 | 0.17 | 0.68 | 1 | 96.68 | 0.55 | 0.46 | 1 | 80.23 | 0.14 | 0.71 |
| GC | 1 | 97.15 | 0.28 | 0.60 | 1 | 96.61 | 3.37 | 0.07 | 1 | 80.24 | 0.13 | 0.72 |
| Informative | 1 | 97.40 | 0.03 | 0.86 | 1 | 95.76 | 1.48 | 0.22 | 1 | 79.11 | 1.26 | 0.26 |
| Length | 1 | 97.31 | 0.11 | 0.74 | 1 | 97.05 | 0.19 | 0.67 | 1 | 79.68 | 0.69 | 0.41 |
| P-distance | 1 | 97.33 | 0.10 | 0.76 | 1 | 96.42 | 0.82 | 0.37 | 1 | 79.35 | 1.02 | 0.31 |
| Polymorphic | 1 | 96.87 | 0.55 | 0.46 | 1 | 97.17 | 0.07 | 0.79 | 1 | 80.07 | 0.30 | 0.58 |
| RI | 1 | 95.60 | 1.83 | 0.18 | 1 | 95.77 | 1.47 | 0.23 | 1 | 79.60 | 0.77 | 0.38 |

### *2.3. Difference of polymorphic sites across codon positions*

A comparison among the fractions of polymorphic sites of the three codon positions indicated that the fraction of codon position 3 is consistently higher than, and significantly different from, that of the other codon positions. Specifically, the fraction of polymorphic sites of codon position 3 was found to be consistently higher than the fraction of polymorphic sites of either codon position 1 or 2 across each of the 78 protein-coding plastome genes under study (Fig. 2). The concomitant evaluation of a significant difference between the fraction means of each codon position using one-way ANOVA and a Kruskal-Wallis test indicated that the fractions of polymorphic sites are significantly different across the three codon positions (ANOVA p-value $< 0.01$; Kruskal-Wallis p-value $< 0.01$). Pairwise comparisons using Wilcoxon rank-sum tests indicated significant differences between mean fraction values when comparisons involved codon position 3 (p-value $< 0.01$ for both such comparisons; Fig. 3).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 May 2019 doi:10.20944/preprints201905.0002.v1



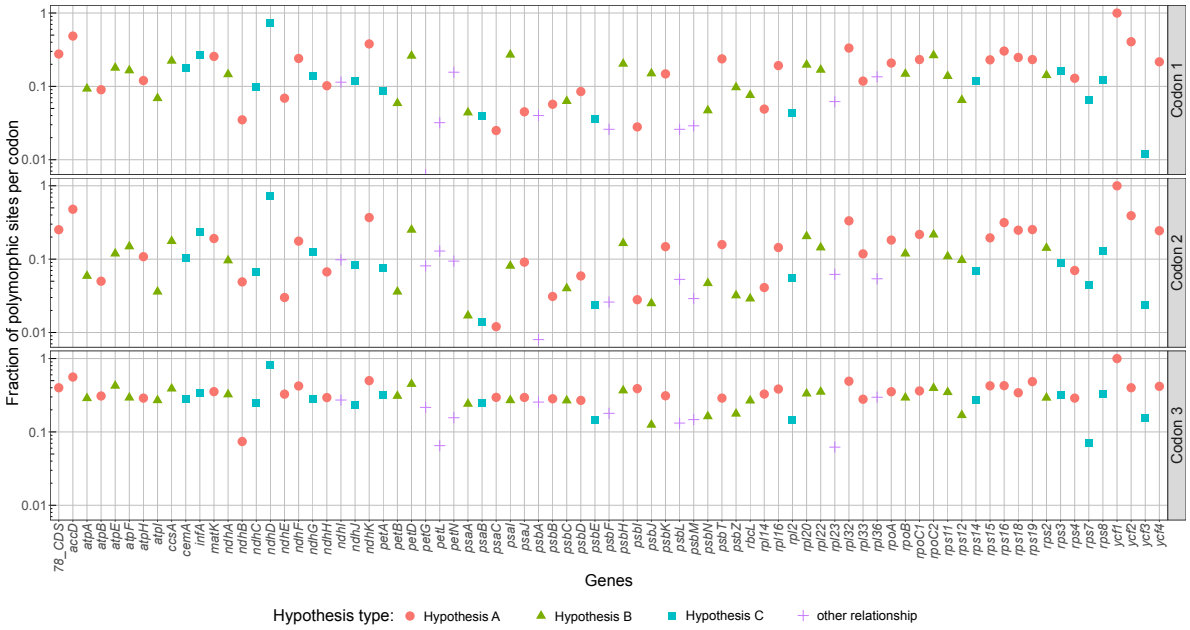


**Figure 2.** Comparison of the fraction of polymorphic sites of the three codon positions across all 78 protein-coding plastome genes under study. The y-axis was transformed logarithmically (logarithm with base 10) to allow for a better visualization of the variability of fraction values within each codon position. Congruence of the phylogenetic relationships inferred under each gene with the phylogenetic hypotheses of Fig. 1 are displayed via differently colored shapes of the individual data points.

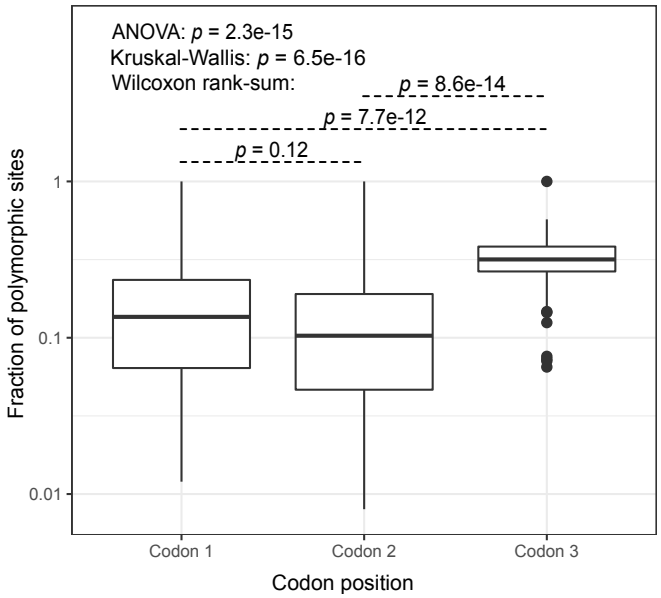


**Figure 3.** Comparison of the mean fraction values of polymorphic sites per codon position across all 78 protein-coding plastome genes. The y-axis was transformed logarithmically (logarithm with base 10) to allow for a better data visualization due to the predominance of fraction values $< 0.5$. P-values of Wilcoxon rank-sum tests to determine significant differences between mean fraction values are displayed on top of the graph.

 doi:10.20944/preprints201905.0002.v1



### 2.4. Incongruence among relationships under different partitioning schemes

Phylogenetic tree inference on the multi-gene alignment recovered incongruent relationships under different data partitioning schemes (Figs. 4, 5). Generally, the phylogenetic reconstructions resulted in highly resolved trees under different data partitioning schemes, with almost all clades exhibiting full branch support. However, the reconstructions displayed incongruence regarding the monophyly of Nymphaeaceae across the different schemes. Reconstructions that were either unpartitioned, partitioned by gene, or partitioned by both gene and codon position recovered relationships consistent with phylogenetic hypothesis A and, thus, supported the monophyly of Nymphaeaceae (Figs. 4a,b,d, 5a,b,d). Reconstructions that were partitioned by codon alone, by contrast, recovered relationships consistent with hypothesis B and, thus, supported the paraphyly of the family (Figs. 4c, 5c). This difference in reconstruction results was seen under both ML (Fig. 4) and BI (Fig. 5) tree inference. Bootstrap and posterior probability support for the nodes determining congruence with either hypothesis was mostly below 80% or 0.80, respectively. Specifically, the recovery of relationships consistent with hypothesis A received low support under partitioning by gene during ML inference (BS=63%; Fig. 4b) but maximum support under unpartitioned analysis and partitioning by gene during BI inference (PP=1.0; Fig. 5a,b). By contrast, the recovery of relationships consistent with hypothesis B received low support under partitioning by codon during ML inference (BS=62%; Fig. 4c) and intermediate support under the same partitioning scheme during BI inference (PP=0.78; Fig. 5c).

### 2.5. Monophyly versus paraphyly of Nymphaeaceae

Based on the phylogenetic reconstructions performed in this investigation, the monophyly of Nymphaeaceae must currently be considered indeterminate. Specifically, the phylogenetic position of the genus *Nuphar* in relation to the rest of Nymphaeaceae and Cabombaceae was found to be dependent on the precise plastome genes analyzed as well as on the data partitioning schemes applied. For example, of the 78 protein-coding plastome genes analyzed, phylogenetic reconstructions on 30 of them support the monophyly of Nymphaeaceae, whereas reconstructions on 38 others support relationships that render Nymphaeaceae paraphyletic (Table 2), often with maximum resampling support (FigureS1). Similarly, reconstructions conducted without data partitioning, under partitioning by gene, and under partitioning by both gene and codon position supported *Nuphar* as the earliest diverging branch within Nymphaeaceae, rendering Nymphaeaceae monophyletic, whereas reconstructions conducted under partitioning by codon consistently recovered *Nuphar* as sister to Cabombaceae, rendering the family paraphyletic (Figs. 4, 5). However, in many of these reconstructions, the position of *Nuphar* was recovered with only intermediate or weak resampling support, making conclusions on the monophyly of Nymphaeaceae speculative.

## 3. Discussion

### 3.1. Phylogenetic incongruence among plastome genes

Considerable phylogenetic incongruence can exist among the genes of the plastid genome, which necessitates customized phylogenetic analyses in plastid phylogenomic studies. Incongruent phylogenetic signal among the genes of the plastome has been reported from an increasing number of investigations [23–26]. Such incongruence can have a considerable impact on phylogeny reconstructions and may even lead to cases where relationships among divergent lineages are driven by only a handful of genes [27]. Given reports of chimeric plastome sequences [28,29] and recombination-based plastid gene conversion [30,31], it appears that the different sections of a plastid genome may not always evolve as the same molecule across time. Instead, positive selection of individual plastome genes [26,32], gene conversion due to recombination-dependent replication [31], and biparental plastome inheritance [33,34] may result in discordant gene phylogenies [28,29].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 May 2019 doi:10.20944/preprints201905.0002.v1



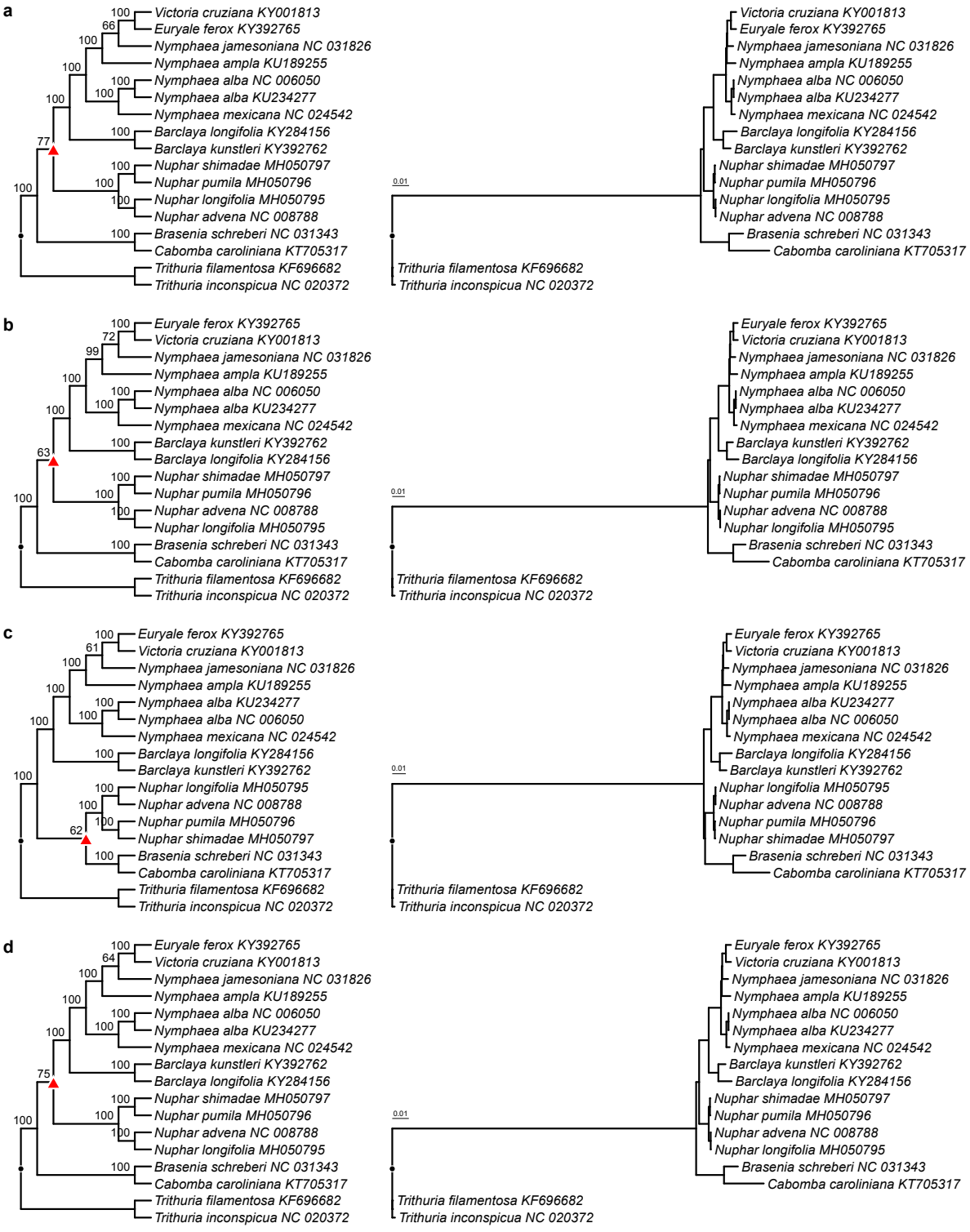


**Figure 4.** Phylogenetic relationships of the 17 taxa of Nymphaeales under study as inferred from the multi-gene alignment of 78 protein-coding plastome genes via ML phylogeny inference under four different data partitioning schemes. The order of the partitioning schemes applied on the sequence matrix are: (**a**) unpartitioned matrix, (**b**) matrix partitioned by gene, (**c**) matrix partitioned by codon, and (**d**) matrix partitioned by gene and codon. The inferred relationships are visualized as cladograms with statistical node support (left) and corresponding phylograms with exact branch lengths (right). Bootstrap support values greater than 50% are given above the branches of each cladogram. The relevant node to determine the monophyly of Nymphaeaceae is highlighted by a red triangle in each cladogram.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 May 2019 doi:10.20944/preprints201905.0002.v1



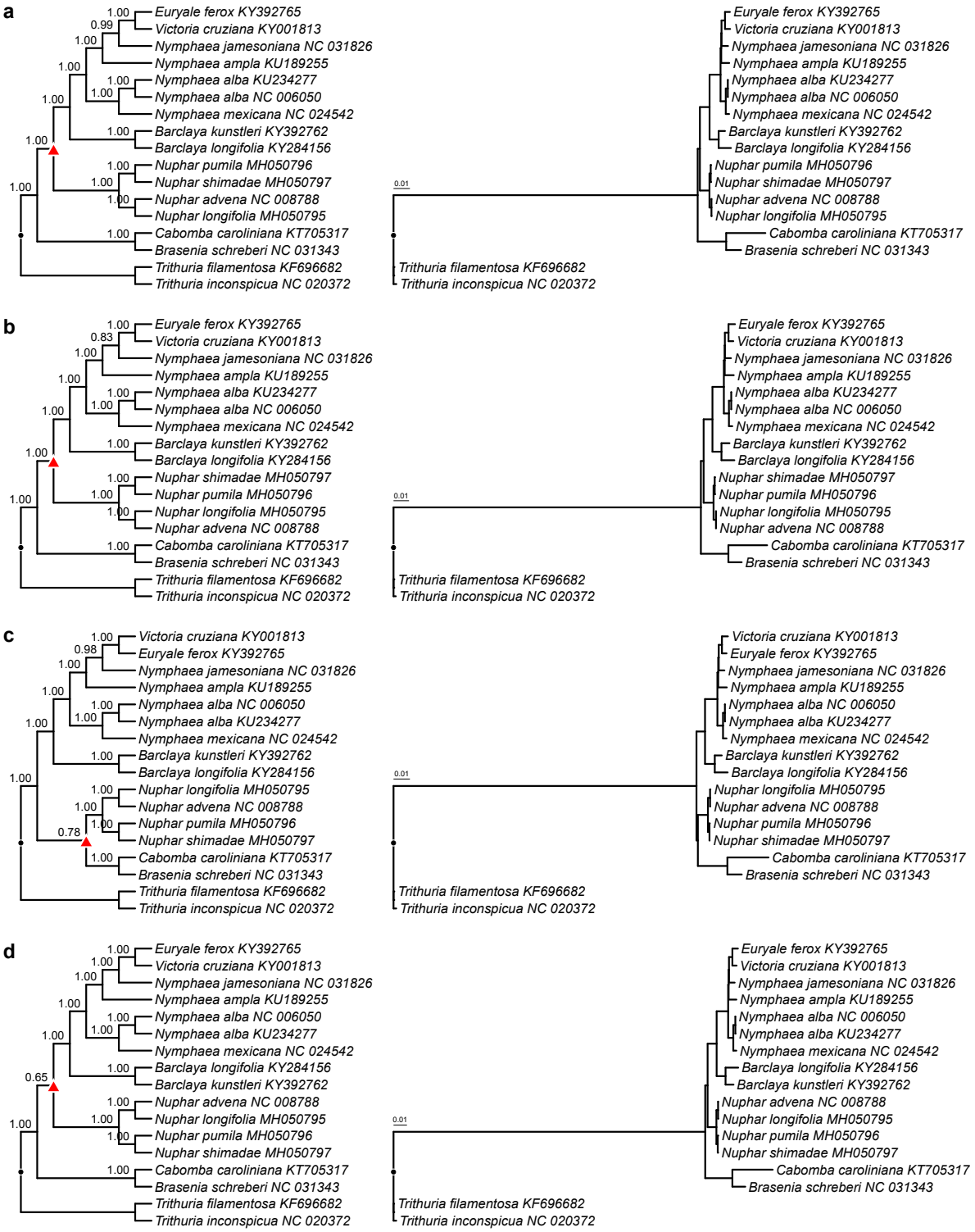


**Figure 5.** Phylogenetic relationships of the 17 taxa of Nymphaeales under study as inferred from the multi-gene alignment of the 78 protein-coding plastome genes via BI phylogeny inference under the four different data partitioning schemes. The layout and settings are as in Fig. 4. Posterior probabilities greater than 0.5 are given above the branches of each cladogram.

According to our results, phylogenetic incongruence among the genes of the plastome affects the phylogenetic placement of *Nuphar* within Nymphaeales. The presence of incongruent phylogenetic signal among plastome genes can unlikely be counteracted by the adjustment of a single parameter but requires the application of fine-scale data partitioning and coalescent methods of phylogeny inference [1,35]. Moreover, future investigations on plastid phylogenomics need to assume the possibility that the individual genes of the plastome do not form a single linkage group (contrary to, for example, [36]) and, thus, do not necessarily result in the same tree topology upon gene-wise phylogeny inference [28,29]. Instead, the divergent evolutionary processes shaping the history of a plastid genome and its individual genes should be appropriately modeled during phylogeny inference [1,37], for example through the application of multispecies coalescent analyses [38].

### *3.2. Phylogenetic incongruence among codon positions*

To evaluate and account for the effect of codon position on phylogenetic inference, data partitioning relative to codon position was applied in the present investigation. Considerable differences in base composition, substitution rates, GC content, and substitutional saturation among codon positions of individual plastome genes have been reported [39,40]. Among the three codon positions, the third position has been found to display particularly high levels of compositional heterogeneity, which can result in aberrant phylogenetic signal [41–43]. Moreover, several investigations reported incongruence between phylogenetic reconstructions based on the signal of only the third codon position compared to reconstructions on the first and second codon position [39,44]. The removal of third codon position data from the alignment of protein-coding genes is, thus, a commonly employed strategy in phylogenetic investigations of early-diverging plant lineages ([45–47], but see [48]). The first and second codon positions are subject to functional constraints of the resulting amino acids and typically display lower rates of substitution and, by extension, homoplasy [42,49]. However, third codon positions often represent a large proportion of the variable sites in most plastome alignments, and their exclusion may result in a drop of phylogenetic signal [39]. Fine-scale data partitioning that assigns different nucleotide substitution models and, thus, separate base and substitution rates to the individual codon positions constitutes an effective solution [35,50]. Hence, data partitioning strategies that model individual codon positions by assuming that nucleotide frequencies and substitutions occur as independent draws from codon-specific probability distributions were employed in the present investigation. Specifically, I treated each codon position independently either (a) across but not within genes (i.e., three partitions, one for each codon position; Figs. 4c, 5c), or (b) both across and within genes (i.e., three times as many partitions as genes analyzed; Figs. 4d, 5d). Such partitioning strategies have been shown to infer phylogenetic relationships of early-diverging angiosperms more reliably than models incognizant of codon position [1,51].

### *3.3. Outgroup selection*

The selection of Hydatellaceae as sole outgroup when reconstructing the phylogenetic relationships of Nymphaeaceae and Cabombaceae represents a suboptimal choice due to the high risk of homoplasy. In their study, He et al. [20] argue that the past uncertainty of intergeneric relationships within Nymphaeaceae may largely have been the results of the selection of inappropriate outgroups. Specifically, they claim that "[t]he proximity of Hydatellaceae to the ingroup, containing Cabombaceae and Nymphaeaceae, makes it a fundamental root in defining character homology in *Nuphar* and the other genera" ([20], p.13). This statement has merit in principle but is challenged by the empirical results of several recent investigations. While it is true that an appropriate outgroup should be closely related to the ingroup, the more meaningful aspect is the low genetic distance that normally exists between closely related groups of species [52]. Taxa that display a high number of autapomorphic changes to the ingroup are poor choices as an outgroup because they increase the rate of homoplasy between in- and outgroup and, by extension, the probability for long-branch attraction [53,54]. The latter scenario appears to be the case when attempting to reconstruct the phylogenetic relationships of

 doi:10.20944/preprints201905.0002.v1



Nymphaeaceae and Cabombaceae while using members of its sister clade Hydatellaceae as outgroup. A high amount of character change exists between the ingroup (Nymphaeaceae and Cabombaceae) and its potential outgroup (Hydatellaceae), as evidenced by the large patristic distance between the two groups [13,14]. In fact, [13] (Fig. 2a therein), [14] (Fig. 6 therein), and [20] (Fig. S2 therein) indicate a considerable amount of nucleotide differences between Nymphaeaceae/Cabombaceae and Hydatellaceae: in the phylogenetic trees reported by each of these investigations, Hydatellaceae are characterized by numerous autapomorphies, as illustrated by long branches between the most recent common ancestor (MRCA) of all Nymphaeales and the MRCA of Trithuria. The use of Hydatellaceae, or rather the currently available plastid genomes thereof, as an outgroup for reconstructing the phylogenetic relationships of Nymphaeaceae is, thus, inadvisable because it likely increases the level of homoplasy during tree inference.

The use of *Amborella* in addition to further taxa of Hydatellaceae as outgroup may be an important aspect in future phylogenomic investigations of Nymphaeaceae and Cabombaceae. The use of only the currently available plastid genomes of Hydatellaceae as outgroup is not advisable for plastid phylogenomic reconstruction of Nymphaeaceae and Cabombaceae because it carries a notable risk of inferring incorrect relationships due to homoplasy. At the same time, few, if any, arguments exist against the selection of the plastid genome of *Amborella* or other early-diverging angiosperms when reconstructing the relationships of this plant lineage. The observation that clade support increases under the selection of Hydatellaceae as outgroup [20] is not a sufficient argument against the selection of *Amborella* or other early-diverging angiosperms because phylogenetic resampling support represents a metric relative to the dataset in use [55]. Graham et al. [54] reported that a randomly-selected outgroup preferentially rooted the ingroup on long terminal ingroup branches. Similarly, [56] and [57] reported zones of parameter space in which outgroups would consistently but incorrectly fail to join with a short branch of an ingroup tree. Future plastid genomic investigations of Nymphaeales should, therefore, aim to include plastid genomes of *Amborella* or other early-diverging angiosperms outside Nymphaeales as well as of Hydatellaceae, provided that additional plastid genomes of Trithuria are generated, as long as they do not display high levels of autapomorphic characters compared to the ingroup. For example, the inclusion of one representative of each of the four sections of the genus Trithuria could be included in future investigations in order minimize the level of homoplasy between the divergent family of Hydatellaceae and the rest of the plant order. Such an inclusion would also benefit the ongoing evaluation of the species diversity of Hydatellaceae, which was recently reported as strongly underestimated [58].

### *3.4. Public data archiving*

Researchers should make all efforts to store the data required to arrive at the conclusions of their investigation in publicly accessible data repositories, along with sufficient details to replicate and assess the published results. The study of [20] provides valuable information on codon usage bias and RNA editing in Nymphaeales. For example, the finding that RNA editing sites exhibit a phylogenetic pattern in which specific editing sites coincide with clades of Nymphaeaceae merits special attention. The study of [20] also addresses the important issue of the monophyly of Nymphaeaceae. However, the validity of the phylogenetic conclusions reached by [20] cannot be assessed directly because their study lacks essential information to replicate their analyses and, by extension, evaluate the results presented. Specifically, the actual set of coding regions extracted from the 17 plastid genomes under study, the precise data partitioning schemes employed, as well as the exact DNA alignments analyzed, were not provided as part of their publication, nor made publicly accessible in common data repositories such as TreeBASE or Dryad (as of 01-Mar-2019). Proper documentation and public archiving of all research results and analysis files constitute one of the primary prerequisites for reproducible and transparent research, particularly in molecular and evolutionary biology [59,60]. In phylogenomic investigations, the concept of meticulous documentation and public archiving of all analysis and

 doi:10.20944/preprints201905.0002.v1



result files is particularly critical because the multi-step bioinformatic analyses that are common in phylogenomic research display a high sensitivity to the precise settings and parameter values employed, often affecting downstream analyses and, ultimately, the final results [61]. To promote the documentation and accessibility of primary research data, many scientific journals have adopted policies that require authors to store all newly generated data in publicly accessible archives as a condition for publication [62,63]. Similarly, recent investigations have devised and recommended procedures to record detailed analysis settings in order to allow other researchers to replicate both the analyses and the presented results, and use them as benchmarks to refine their own analyses [60,61]. The overall objective of these policies and guidelines is to ensure the widest possible accessibility of all relevant data of a published study and to promote re-evaluation, re-usability, and re-purposing of that data [64]. In the spirit of this objective, all research data generated and analyzed as part of this investigation has been publicly archived at, and made available for re-analysis from, Zenodo, including all DNA alignments, data partition schemes, tree inference files, and phylogenetic trees analyzed.

## 4. Materials and Methods

### *4.1. Assembly of original dataset*

The assembly of the original dataset of [20] was not possible based on the published information alone; only email correspondence with the authors provided the necessary information on the identity of the protein-coding genes under study. To re-assess the phylogenetic conclusions presented in [20], I aimed to assemble the same dataset as analyzed in their study by extracting the set of 66 protein-coding genes that the authors cited as being the basis for their phylogenetic analyses. Specifically, the authors stated that the "Maximum Likelihood (ML) and Bayesian Inference (BI) phylogenetic tree was based on 66 protein codon [sic!] genes" ([20], p.10). This number of genes is notably smaller than the 79 unique protein-coding genes that [13], as well as [20] themselves, reported as the full gene complement for the plastid genomes of Nymphaeaceae. Unfortunately, no indication as to the precise identity of these 66 genes is provided in their publication. Thus, the task of assembling the same dataset as analyzed by [20] was impossible to conduct without requesting additional information from the authors. Consequently, the first and corresponding authors of [20] were contacted to provide information on the identity of the protein-coding genes used in their study. The ensuing email correspondence with one of the first authors on 25-Jan-2019 and 26-Jan-2019 initially informed me that not 66, but a total of 73 protein-coding genes had been extracted from the plastid genomes and concatenated for their phylogenetic analyses. A set of seven protein-coding genes present in all plastid genomes of Nymphaeaceae would, thus, not have been included in their original analyses. Upon a request for clarification, the earlier information was revised, and I was informed that a total of 78 protein-coding genes had been extracted from the plastid genomes and concatenated for phylogenetic analysis. Thus, all but one protein-coding gene present in the full plastome gene complement of Nymphaeales (i.e., *clpP*) would have been included in the original analyses of [20]. For the present investigation, this latest clarification by the authors of [20] was assumed to be correct and used for the assembly of the original dataset.

### *4.2. Manual correction of annotations*

Upon assembly of the original dataset of [20], genes were evaluated and corrected for potentially incorrect annotations. The annotations of four different protein-coding genes of the plastid genomes of *Barclaya kunstleri* (GenBank accession number KY392762) and *Euryale ferox* (KY392765) were identified to be potentially incorrect. For example, the annotations of the genes *atpH* and *rpl22* in the plastid genome of *Barclaya kunstleri* displayed internal stop codons that should have marked the end of their amino acid sequences; the coding sequences of these two genes were 21 nucleotides (i.e., 7 amino acids) and 9 nucleotides (i.e., 3 amino acids), respectively, longer than permitted under a default reading frame. Similarly, the annotations of the genes *atpA* and *rpoA* in the plastid genome of *Euryale ferox*

 doi:10.20944/preprints201905.0002.v1



displayed internal stop codons that should have marked the end of their amino acid sequences; the two genes were 9 nucleotides (i.e., 3 amino acids) and 21 nucleotides (i.e., 7 amino acids) too long, respectively. None of these cases is likely the result of RNA editing, as the particular stop codon positions do not occur in any other plastid genome under study, nor were they included by [20] in the list of nucleotide positions affected by RNA editing. Each of these presumably incorrect annotations was, thus, corrected manually using Geneious v.10.2.3 [65] prior to extracting their DNA sequences for gene-wise DNA alignment. Without such corrections, the gene-wise DNA alignments, as well as the subsequent phylogenetic analyses, would carry the risk of being biased due to an incorrect length of the extracted sequences.

### *4.3. DNA sequence extraction and alignment*

All protein-coding genes were extracted and aligned on a gene-by-gene basis using an automated alignment procedure. Each of the 78 protein-coding genes analyzed by [20] was bioinformatically extracted from existing GenBank records and then aligned based on their amino acid sequences using a two-step procedure. Both steps were automated by script 9 of the pipeline of [61]. First, the genes were extracted from the most recent GenBank records of the plastid genomes of the 17 species under study. The hypothetical plastome genes *ycf15* and *ycf68* as well as the plastome open reading frames *orf42* and *orf56* were not extracted due to their uncertain gene status (reviewed in [13]). Similarly, gene *clpP* was not extracted, as this gene had not been included in the analyses of [20]. Upon extraction, the sequences were grouped by gene name, translated from nucleotides to amino acids, aligned gene by gene based on their amino acid sequences using MAFFT v.7.309 [66] under default settings (gap open penalty of 1.23, automatic determination of sequence direction and best alignment algorithm), back-translated to nucleotide sequences, and saved as individual gene alignments. Thus, a total of 78 DNA sequence alignments, each representing a different protein-coding gene, was saved for subsequent phylogenetic analyses. At the same time, a multi-gene alignment was generated by concatenating the individual gene alignments in the same gene order as found in the actual plastid genomes, which is possible due to the conserved gene synteny among the plastid genomes of Nymphaeales [61]. All 78 gene-wise DNA alignments, as well as the concatenated multi-gene alignment, have been deposited to, and made available for re-analysis via, Zenodo (https://zenodo.org/record/2613673).

Due to small methodological differences in the alignment process, the multi-gene DNA alignment analyzed in this investigation may be different from its counterpart in [20]. The present investigation is applying an alignment procedure that has small advantages over more commonly employed plastid phylogenomic alignment procedures and may, thus, result in a slightly different DNA sequence alignment than the one generated and analyzed by [20]. However, since the alignment of [20] has not been made publicly accessible as part of their publication, it is not possible to make firm statements regarding such differences. The application of particularly two methodological aspects may have resulted in different alignments: (a) In the present investigation, the DNA sequences were automatically aligned before concatenating the individual genes. Gene-wise alignments of phylogenomic data can be a time-intensive and error-prone process unless automated [67]. Automated gene-by-gene alignment ensures that any insertion or deletion at the 5′ or 3′ ends of a coding region does not result in an overlap of different genes in the resulting multi-gene alignment, which would not be the case when aligning the genes after their concatenation. (b) In the present investigation, the DNA sequences were aligned based on their amino acid, not on their nucleotide sequences, which some contemporary plastid phylogenomic studies continue to do (e.g., [68,69]). Conducting the alignment on amino acid instead of nucleotide sequences ensures that the trinucleotide reading frame of codons is preserved and, thus, insertions/deletions are always a multiple of three. Both methodological aspects represent beneficial policies in plastid phylogenomic alignment procedures, especially if automated bioinformatically [61].

 doi:10.20944/preprints201905.0002.v1



*4.4. Sequence variability and homoplasy metrics*

For each alignment under study, a series of sequence variability and homoplasy metrics was calculated to explore possible effects on the phylogenetic trees inferred from the different plastome genes. Indices of sequence variability, substitution rates, and homoplasy are important measures to characterize genomic regions [70,71]. In the present investigation, a total of eleven sequence variability and homoplasy metrics was calculated for each of the 78 gene-wise DNA alignments under study. The selected indices of sequence variability are quantitative metrics and aim to characterize the magnitude of the fluctuations among aligned DNA sequences [72,73]; they are: the fraction of polymorphic sites in the alignment, the fraction of parsimony-informative sites in the alignment, and the maximum uncorrected p-distance between sequences of the alignment. The selected indices of homoplasy are also quantitative metrics and aim to characterize the magnitude of the homoplastic character changes across aligned DNA sequences [74]; they are: the consistency index [75], the retention index [76], and the rescaled consistency index [76]. Moreover, I calculated the fraction of polymorphic sites for each of the three codon positions in the alignments. Differential substitution rates among the three codon positions and an effect on phylogeny inference have been reported in several investigations [77,78], including those of plastid genome evolution among land plants [39,43]. Each of these sequence variability and homoplasy metrics may represent a potential explanatory variable for the incongruent phylogenetic signal between the different plastome genes of Nymphaeales. It is, thus, important to evaluate if the results of gene-wise phylogenetic inference correlate with any of the calculated sequence variability and homoplasy metrics at a statistically significant level. All sequence variability and homoplasy metrics were inferred in R v.3.5.1 [79] using the R packages ape v.5.2 [80], ips v.0.0-7 [81], and phangorn v.2.4.0 [82].

*4.5. DNA model selection*

For each DNA sequence alignment under study, the best-fitting nucleotide substitution model was inferred in R. Specifically, for each of the 78 gene-wise alignments as well as for the multi-gene alignment, R package phangorn was employed to compare the fit of different nucleotide substitution models to the sequence data. The sample-size corrected Akaike information criterion [83] was used as the selection criterion for best fit among the tested models. The best-fitting nucleotide substitution model identified for each alignment is given in TableS1.

*4.6. Phylogenetic inference*

To re-assess the phylogenetic conclusions of [20], the evolutionary relationships of the study taxa were reconstructed under the gene-wise alignments and the multi-gene alignment. Multiple investigations have reported incongruence between different genes or sections of the plastid genome [24,26]. The preliminary analyses of this investigation indicated a similar incongruence (Table 1). Hence, the phylogenetic congruence between individual genes of Nymphaeales plastid genomes was evaluated by reconstructing the phylogenetic relationships among the 17 study taxa under each of the 78 individual gene alignments as well as the multi-gene alignment. Phylogenetic inference under the individual gene alignments was conducted via the maximum likelihood (ML) inference criterion in R. Specifically, script 11 of the pipeline of [61] was employed to conduct phylogenetic tree inference under ML, including the evaluation of branch support via the calculation of 100 bootstrap (BS) replicates. To render the best ML trees of different genes visually comparable, the phylogram with the highest likelihood score of each gene was converted to an ultrametric tree via the penalized likelihood approach of the R package ape and a root age of 1. Phylogenetic inference under the multi-gene alignment was conducted via ML as well as the Bayesian inference (BI) criterion. Inference via ML was conducted with RAxML v.8.2.9 [84] using the thorough ML optimization option. Branch support under ML was calculated via 1,000 BS replicates using the rapid BS algorithm [85]. Inference via BI was conducted with MrBayes v.3.2.5 [86] using two parallel Markov Chain Monte Carlo (MCMC)

 doi:10.20944/preprints201905.0002.v1



runs for a total 20 million generations. The initial 75% of all MCMC trees were discarded as burn-in, and post-burn-in trees were summarized as a majority rule consensus tree, with branch support given as posterior probability (PP) values. Best-fitting nucleotide substitution models were specified for phylogenetic inference under individual gene alignments; specifically, the best-fitting model of the gene under study was employed. The post-burnin posterior tree distributions and the MCC trees of the phylogenetic analyses under BI as well as the trees with the highest likelihood scores and the bootstrap replicate trees of the phylogenetic analyses under ML have been deposited to, and made available for re-analysis from, Zenodo.

### *4.7. Data partitioning*

Four different partitioning strategies were applied during phylogenetic inference to evaluate the impact of data partitioning as well as of the individual codon positions on the phylogenetic conclusions. An effect of different partitioning schemes on the phylogenetic reconstruction of Nymphaeales has been reported by [13]. Similarly, a measurable impact of different codon positions on the phylogenetic reconstruction of basal angiosperms was reported by citeYangEtAl2007. Both studies highlight the importance of data partitioning in plastid phylogenomic analyses of early-diverging flowering plants. To explore the dependence of the phylogenetic conclusions reached here and by [20] to data partitioning, four different partitioning strategies were applied during phylogeny reconstruction. First, phylogenetic analyses were conducted on an unpartitioned matrix in which the entire multi-gene DNA alignment was analyzed under the nucleotide substitution model GTR+I+G; a single partition was analyzed under this strategy. Second, phylogenetic analyses were conducted on a partitioned matrix in which each of the 78 protein-coding genes was analyzed under its best-fitting nucleotide substitution model; a total of 78 partitions were analyzed under this strategy. Third, phylogenetic analyses were conducted on a partitioned matrix in which each of the three codon positions across the alignment was grouped into a separate partition; a total of three partitions were analyzed under this strategy. Fourth, phylogenetic analyses were conducted on a partitioned matrix in which each of the 78 protein-coding genes, as well as each of the three codon positions across these genes, were grouped into separate partitions; a total of 234 partitions was analyzed under this strategy. Each of these partitioning strategies was applied in phylogenetic reconstructions via ML and BI. Taken together, this diverse set of partitioning schemes allows the evaluation of phylogenetic congruence among and across individual genes of the plastid genome while simultaneously accounting for the effect of codon position on phylogenetic inference. The input files for tree inference which specify each of the four data partitioning schemes analyzed, have been deposited to, and made available for re-analysis from, Zenodo.

### *4.8. Effects of sequence variability metrics on phylogenetic inference*

Logistic regression models were employed to explore potential effects of the sequence variability and homoplasy metrics inferred for each gene alignment on the relationships recovered under gene-wise phylogeny inference. To assess and identify shared factors among the genes that encode for specific phylogenetic relationships of Nymphaeaceae (i.e., hypotheses A, B or C in Fig. 1), a series of multiple logistic regression models were defined and their fit to the data evaluated. Specifically, I tested for statistically significant effects of the sequence variability/homoplasy metrics calculated from the individual gene alignments on the phylogenetic relationships inferred from the same genes. The recovered relationships constitute the dependent variables of the models, whereas the sequence variability/homoplasy metrics constitute the independent variables. The models defined comprised a null and several alternative models. The null model specified no effects by any of the sequence variability/homoplasy metrics, whereas the most complex of the alternative models specified effects by all of them. Models more complex than the null model were generated automatically through the stepwise addition of individual sequence variability/homoplasy metrics. Likelihood ratio tests (LRTs) and the Akaike information criterion (AIC) [87] were used as selection criteria for identifying best model fit in pairwise comparisons among the specified models. A significance threshold of $\alpha$

 doi:10.20944/preprints201905.0002.v1



= 0.01 was defined for evaluating the significance of the likelihood ratio statistic in relation to a $\chi^2$ distribution. This setup allowed me to evaluate if any of the sequence variability/homoplasy metrics correlated with specific relationships recovered under gene-wise phylogeny reconstructions and, by extension, if any of the metrics represented an explanatory variable for the incongruent phylogenetic signal between the plastome genes. All tests and model specifications were conducted in R using the in-built functions on generalized linear models and the R package car v.3.0 [88].

### *4.9. Comparison of polymorphic sites across codon positions*

To assess the variability within, and the difference between, the fraction of polymorphic sites of the three codon positions, two different analyses/visualizations were conducted. First, the variability of the fraction of polymorphic sites of each codon position was visualized across all 78 protein-coding plastome genes in R using the R package ggplot2 v.3.1.0 [89]. This visualization allows the determination if one of the three fractions is consistently higher than the other two. Second, the difference between the three codon positions with regard to the fraction of polymorphic sites was evaluated via pairwise comparisons of their mean fraction values using Wilcoxon rank-sum tests and via overall comparisons using one-way analysis of variance (ANOVA) as parametric and a Kruskal-Wallis test as non-parametric statistical tests. These tests were conducted in R using the in-built statistics functions and the R package ggpubr v.0.2 [90]. A significance threshold of $\alpha = 0.01$ was defined for evaluating the significance of the mean comparisons.

## 5. Conclusions

The present investigation aimed to re-assess the conclusion of monophyly of Nymphaeaceae as presented by [20] by re-analyzing their dataset and performing additional phylogenetic reconstructions and multivariate statistical analyses. Specifically, I aimed to assess if the plastid phylogenomic reconstructions were congruent across different plastome genes and different data partitioning schemes, and which factors might be associated with potential cases of incongruence. Given the results of the present investigation, the conclusion by [20] that the phylogenetic position of *Nuphar* had now been ascertained and the monophyly of Nymphaeaceae resolved with convincing statistical support cannot be corroborated. Instead, considerable incongruence among the phylogenetic relationships of Nymphaeales inferred from different plastome genes and under different data partitioning schemes was identified. In fact, the results indicate that the phylogenetic signal among different plastome genes and the variability among the codon positions of these genes is strongly heterogeneous, and that even fine-scale data partitioning is insufficient to account for the conflicting phylogenetic signal. Furthermore, the results of this investigation suggest that the previous recovery of Nymphaeaceae as monophyletic [20] may have been primarily supported by the highly dynamic and possibly homoplastic third codon positions of the protein-coding plastome genes. Future plastid phylogenomic investigations on the evolutionary history of Nymphaeales should aim to evaluate and compare the precise phylogenetic signal among the coding regions under study. Likewise, a more extensive taxon sampling is advisable for future investigations of Nymphaeales, in particular with regard to species of the family Hydatellaceae which may a considerable influence on the phylogenetic placement of the genus *Nuphar* and, by extension, the inference of monophyly of Nymphaeaceae due to their long patristic distance between stem and crown node. Finally, future investigations should make all efforts to archive the data required to assess their analyses and conclusions in sufficient detail in publicly accessible repositories. In summary, I concur with the final conclusion of [20] that additional research on the phylogenetic history of Nymphaeales is needed.

**Supplementary Materials:** The following are available online at http://www.mdpi.com//xx/1/5/s1. Figure S1: Results of the gene-wise phylogenetic reconstructions under the ML inference criterion on each of the 78 protein-coding plastome genes under study. Bootstrap support values are given above the branches. For easier viewing, all taxa of Nymphaeaceae are indicated by a red bar to the right of their taxon names. Table S1: Results of calculating the fractions of polymorphic sites of the three codon positions across the 78 protein-coding plastome genes under study that form the basis for the multiple logistic regression tests of possible effects on the

phylogenetic tree inference as well as the overall and pairwise comparisons of the mean fraction values using parametric and non-parametric statistical tests. Also displayed are the best-fitting nucleotide substitution models identified for each of the plastome genes under study. Abbreviations are used as in Tables 1, 2 and 3. Table S2: Results of calculating eight different sequence variability and homoplasy metrics that form the basis for the multiple logistic regression tests of possible effects on the phylogenetic tree inference. Abbreviations are used as in Tables 1, 2 and 3.

**Author Contributions:** M.G. conceived this investigation, conducted all analyses, wrote the manuscript, and generated all figures, tables and supplemental files.

**Funding:** This investigation was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 418670221 – and by a start-up grant of the Freie Universität Berlin (Initiativmittel der Forschungskommission), both to M.G.

**Acknowledgments:** The author wishes to thank Prof. Thomas Borsch of the Freie Universität Berlin and the Botanischer Garten und Botanisches Museum Berlin for valuable discussions on the evolutionary history of Nymphaeaceae. The author also wishes to thank Prof. Robert K. Jansen of the University of Texas at Austin for valuable discussions on gene-wise plastid phylogenomic analyses. Furthermore, the author thanks Nicholas Turland of the Botanischer Garten und Botanisches Museum Berlin as well as Prof. Tod F. Stuessy of the Ohio State University for valuable feedback on a preliminary version of this manuscript. The author also acknowledges the high-performance computing service of the ZEDAT of the Freie Universität Berlin for providing allocations of computing time.

**Conflicts of Interest:** The author declares no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| AIC | Akaike information criterion |
| ANOVA | analysis of variance |
| BI | Bayesian inference |
| BS | bootstrap |
| LRT | likelihood ratio test |
| MCMC | Markov Chain Monte Carlo |
| ML | maximum likelihood |
| MRCA | most recent common ancestor |
| PP | posterior probability |

## References

1. Zhong, B.; Betancur-R, R. Expanded taxonomic sampling coupled with gene genealogy interrogation provides unambiguous resolution. *Genome Biology and Evolution* **2017**, *9*, 3154–3161. doi:10.1093/gbe/evx233.
2. Gitzendanner, M.A.; Soltis, P.S.; Wong, G.K.S.; Ruhfel, B.R.; Soltis, D.E. Plastid phylogenomic analysis of green plants: A billion years of evolutionary history. *American Journal of Botany* **2018**, *105*, 291–301. doi:10.1002/ajb2.1048.
3. Ran, J.H.; Shen, T.T.; Wang, M.M.; Wang, X.Q. Phylogenomics resolves the deep phylogeny of seed plants and indicates partial convergent or homoplastic evolution between Gnetales and angiosperms. *Proceedings of the Royal Society B: Biological Sciences* **2018**, *285*, 20181012. doi:10.1098/rspb.2018.1012.
4. Borsch, T.; Loehne, C.; Wiersema, J. Phylogeny and evolutionary patterns in Nymphaeales: Integrating genes, genomes and morphology. *Taxon* **2008**, *57*, 1052–1081. doi:10.2307/27756765.
5. Loehne, C.; Wiersema, J.H.; Borsch, T. The unusual Ondinea, actually just another Australian water-lily of Nymphaea subg. Anecphya (Nymphaeaceae). *Willdenowia* **2009**, *39*, 55–58. doi:10.3372/wi.39.39104.
6. Vialette-Guiraud, A.C.M.; Alaux, M.; Legeai, F.; Finet, C.; Chambrier, P.; Brown, S.C.; Chauvet, A.; Magdalena, C.; Rudall, P.J.; Scutt, C.P. Cabomba as a model for studies of early angiosperm evolution. *Annals of Botany* **2011**, *108*, 589–598. doi:10.1093/aob/mcr088.
7. De Lima, C.; Dos Santos, F.d.A.R.; Giulietti, A.M. Morphological strategies of Cabomba (Cabombaceae), a genus of aquatic plants. *Acta Botanica Brasilica* **2014**, *28*, 327–338. doi:10.1590/0102-33062014abb3439.

8. Sokoloff, D.D.; Remizowa, M.V.; Macfarlane, T.D.; Rudall, P.J. Classification of the early-divergent angiosperm family Hydatellaceae: One genus instead of two, four new species and sexual dimorphism in dioecious taxa. *Taxon* **2008**, *57*, 179–200.
9. Iles, W.J.D.; Rudall, P.J.; Sokoloff, D.D.; Remizowa, M.V.; Macfarlane, T.D.; Logacheva, M.D.; Graham, S.W. Molecular phylogenetics of Hydatellaceae (Nymphaeales): sexual-system homoplasy and a new sectional classification. *American Journal of Botany* **2012**, *99*, 663–676. doi:10.3732/ajb.1100524.
10. Borsch, T.; Loehne, C.; Mbaye, M.S.; Wiersema, J.H. Towards a complete species tree of Nymphaea: Shedding further light on subg. Brachyceras and its relationships to the Australian water-lilies. *Telopea* **2011**, *13*, 193–217.
11. Saarela, J.M.; Rai, H.S.; Doyle, J.A.; Endress, P.K.; Mathews, S.; Marchant, A.D.; Briggs, B.G.; Graham, S.W. Hydatellaceae identified as a new branch near the base of the angiosperm phylogenetic tree. *Nature* **2007**, *446*, 5–8. doi:10.1038/nature05612.
12. Friedman, W.E. Hydatellaceae are water lilies with gymnospermous tendencies. *Nature* **2008**, *453*, 94–97. doi:10.1038/nature06733.
13. Gruenstaeudl, M.; Nauheimer, L.; Borsch, T. Plastid genome structure and phylogenomics of Nymphaeales: conserved gene order and new insights into relationships. *Plant Systematics and Evolution* **2017**, *303*, 1251–1270. doi:10.1007/s00606-017-1436-5.
14. Li, B.; Zheng, Y. Dynamic evolution and phylogenomic analysis of the chloroplast genome in Schisandraceae. *Scientific Reports* **2018**, *8*, 9285. doi:10.1038/s41598-018-27453-7.
15. Loehne, C.; Borsch, T.; Wiersema, J.H. Phylogenetic analysis of Nymphaeales using fast-evolving and noncoding chloroplast markers. *Botanical Journal of the Linnean Society* **2007**, *154*, 141–163. doi:10.1111/j.1095-8339.2007.00659.x.
16. Biswal, D.K.; Debnath, M.; Kumar, S.; Tandon, P. Phylogenetic reconstruction in the order Nymphaeales: ITS2 secondary structure analysis and in silico testing of maturase K (matK) as a potential marker for DNA barcoding. *BMC Bioinformatics* **2012**, *13*, S26. doi:10.1186/1471-2105-13-S17-S26.
17. Yoo, M.J.; Soltis, P.S.; Soltis, D.E. Expression of floral MADS-box genes in two divergent water lilies: Nymphaeales and Nelumbo. *International Journal of Plant Sciences* **2010**, *171*, 121–146. doi:10.1086/648986.
18. Taylor, M.L.; Cooper, R.L.; Schneider, E.L.; Osborn, J.M. Pollen structure and development in Nymphaeales: Insights into character evolution in an ancient angiosperm lineage. *American Journal of Botany* **2015**, *102*, 1685–1702. doi:10.3732/ajb.1500249.
19. Chen, F.; Liu, X.; Yu, C.; Chen, Y.; Tang, H.; Zhang, L. Water lilies as emerging models for Darwin's abominable mystery. *Horticulture Reasearch* **2017**, *4*, 17051. doi:10.1038/hortres.2017.51.
20. He, D.; Gichira, A.W.; Li, Z.; Nzei, J.M.; Guo, Y.; Wang, Q.; Chen, J. Intergeneric relationships within the early-diverging angiosperm family Nymphaeaceae based on chloroplast phylogenomics. *International Journal of Molecular Sciences* **2018**, pp. 9–11. doi:10.3390/ijms19123780.
21. Piel, W.; Donoghue, M.; Sanderson, M. TreeBASE: A database of phylogenetic knowledge. In *To the interoperable 'Catalog of Life' with partners - Species 2000 Asia Oceania*; Shimura, J.; Wilson, K.; Gordon, D., Eds.; National Institute for Environmental Studies: Ibaraki, Japan, 2002; chapter 8, pp. 41–47.
22. Greenberg, J.; White, H.C.; Carrier, S.; Scherle, R. A metadata best practice for a scientific data repository. *Journal of Library Metadata* **2009**, *9*, 194–212. doi:10.1080/19386380903405090.
23. Parks, M.; Cronn, R.; Liston, A. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biology* **2009**, *7*, 84. doi:10.1186/1741-7007-7-84.
24. Barrett, C.F.; Specht, C.D.; Leebens-Mack, J.; Stevenson, D.W.; Zomlefer, W.B.; Davis, J.I. Resolving ancient radiations: Can complete plastid gene sets elucidate deep relationships among the tropical gingers (Zingiberales)? *Annals of Botany* **2014**, *113*, 119–133. doi:10.1093/aob/mct264.
25. Li, Y.; Yang, Y.; Yu, L.; Du, X.; Ren, G. Plastomes of nine hornbeams and phylogenetic implications. *Ecology and Evolution* **2018**, *8*, 8770–8778. doi:10.1002/ece3.4414.
26. Saarela, J.M.; Burke, S.V.; Wysocki, W.P.; Barrett, M.D.; Clark, L.G.; Craine, J.M.; Peterson, P.M.; Soreng, R.J.; Vorontsova, M.S.; Duvall, M.R. A 250 plastome phylogeny of the grass family (Poaceae): Topological support under different data partitions. *PeerJ* **2018**, *6*, e4299. doi:10.7717/peerj.4299.
27. Shen, X.X.; Hittinger, C.T.; Rokas, A. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology and Evolution* **2017**, *1*, 0126. doi:10.1038/s41559-017-0126.

28. Sullivan, A.R.; Schiffthaler, B.; Thompson, S.L.; Street, N.R.; Wang, X.R. Interspecific plastome recombination reflects ancient reticulate evolution in Picea (Pinaceae). *Molecular Biology and Evolution* **2017**, *34*, 1689–1701. doi:10.1093/molbev/msx111.
29. Zhu, A.; Fan, W.; Adams, R.P.; Mower, J.P. Phylogenomic evidence for ancient recombination between plastid genomes of the Cupressus-Juniperus-Xanthocyparis complex (Cupressaceae). *BMC Evolutionary Biology* **2018**, *18*, 137. doi:10.1186/s12862-018-1258-2.
30. Marechal, A.; Brisson, N. Recombination and the maintenance of plant organelle genome stability. *New Phytologist* **2010**, *186*, 299–317.
31. Ruhlman, T.A.; Zhang, J.; Blazier, J.C.; Sabir, J.S.M.; Jansen, R.K. Recombination-dependent replication and gene conversion homogenize repeat sequences and diversify plastid genome structure. *American Journal of Botany* **2017**, *104*, 559–572. doi:10.3732/ajb.1600453.
32. Carbonell-Caballero, J.; Alonso, R.; Iba, V.; Terol, J.; Talon, M.; Dopazo, J. A phylogenetic analysis of 34 chloroplast genomes elucidates the relationships between wild and domestic species within the genus Citrus. *Molecular Biology and Evolution* **2015**, *32*, 2015–2035. doi:10.1093/molbev/msv082.
33. Wolfe, A.D.; Randle, C.P. Recombination, heteroplasmy, haplotype polymorphism, and paralogy in plastid genes: Implications for plant molecular systematics. *Systematic Botany* **2004**, *29*, 1011–1020.
34. Zhang, Q.; Sodmergen. Why does biparental plastid inheritance revive in angiosperms? *Journal of Plant Research* **2010**, *123*, 201–206. doi:10.1007/s10265-009-0291-z.
35. Kainer, D.; Lanfear, R. The effects of partitioning on phylogenetic inference. *Molecular Biology and Evolution* **2015**, *32*, 1611–1627. doi:10.1093/molbev/msv026.
36. Lu, L.; Cox, C.J.; Mathews, S.; Wang, W.; Wen, J.; Chen, Z. Optimal data partitioning, multispecies coalescent and Bayesian concordance analyses resolve early divergences of the grape family. *Cladistics* **2018**, *34*, 57–77. doi:10.1111/cla.12191.
37. Arenas, M. Trends in substitution models of molecular evolution. *Frontiers in Genetics* **2015**, *6*, 319. doi:10.3389/fgene.2015.00319.
38. Bernhardt, N.; Brassac, J.; Kilian, B.; Blattner, F.R. Dated tribe-wide whole chloroplast genome phylogeny indicates recurrent hybridizations within Triticeae. *BMC Evolutionary Biology* **2017**, *17*, 141. doi:10.1186/s12862-017-0989-9.
39. Ruhfel, B.R.; Gitzendanner, M.A.; Soltis, P.S.; Soltis, D.E.; Burleigh, J. From algae to angiosperms-inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evolutionary Biology* **2014**, *14*, 23. doi:10.1186/1471-2148-14-23.
40. Suzuki, H.; Morton, B.R. Codon adaptation of plastid genes. *PLOS ONE* **2016**, *11*, e0154306. doi:10.1371/journal.pone.0154306.
41. Wolf, P.G.; Karol, K.G. Plastomes of bryophytes, lycophytes and ferns. In *Genomics of chloroplasts and mitochondria*; Bock, R.; Knoop, V., Eds.; Springer: Dordrecht, Netherlands, 2012; chapter 4, pp. 89–102. doi:10.1007/978-94-007-2920-9_4.
42. Cox, C.J.; Li, B.; Foster, P.G.; Embley, T.M.; Civan, P. Conflicting phylogenies for early land plants are caused by composition biases among synonymous substitutions. *Systematic Biology* **2014**, *63*, 272–279. doi:10.1093/sysbio/syt109.
43. Fang, L.; Leliaert, F.; Zhang, Z.H.; Penny, D.; Zhong, B.J. Evolution of the Chlorophyta: Insights from chloroplast phylogenomic analyses. *Journal of Systematics and Evolution* **2017**, *55*, 322–332. doi:10.1111/jse.12248.
44. Yang, Y.; Zhu, J.; Feng, L.; Zhou, T.; Bai, G.; Yang, J.; Zhao, G. Plastid genome comparative and phylogenetic analyses of the key genera in Fagaceae: Highlighting the effect of codon composition bias in phylogenetic inference. *Frontiers in Plant Science* **2018**, *9*, 82. doi:10.3389/fpls.2018.00082.
45. Goremykin, V.V.; Viola, R.; Hellwig, F.H. Removal of noisy characters from chloroplast genome-scale data suggests revision of phylogenetic placements of Amborella and Ceratophyllum. *Journal of Molecular Evolution* **2009**, *68*, 197–204. doi:10.1007/s00239-009-9206-9.
46. Fucikova, K.; Lewis, P.O.; Lewis, L.A. Chloroplast phylogenomic data from the green algal order Sphaeropleales (Chlorophyceae, Chlorophyta) reveal complex patterns of sequence evolution. *Molecular Phylogenetics and Evolution* **2016**, *98*, 176–183. doi:10.1016/j.ympev.2016.01.022.

47. Jackson, C.; Knoll, A.H.; Chan, C.X.; Verbruggen, H. Plastid phylogenomics with broad taxon sampling further elucidates the distinct evolutionary origins and timing of secondary green plastids. *Scientific Reports* **2018**, *8*, 1523. doi:10.1038/s41598-017-18805-w.
48. Yang, X.; Tuskan, G.A.; Tschaplinski, T.J.; Cheng, M.Z.M. Third-codon transversion rate-based Nymphaea basal angiosperm phylogeny - concordance with developmental evidence. *Nature Precedings* **2007**. doi:10.1038/npre.2007.320.1.
49. Townsend, J.P. Profiling phylogenetic informativeness. *Systematic Biology* **2007**, *56*, 222–231. doi:10.1080/10635150701311362.
50. Lanfear, R.; Calcott, B.; Ho, S.Y.W.; Guindon, S. PartitionFinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution* **2012**, *29*, 1695–1701. doi:10.1093/molbev/mss020.
51. Goremykin, V.V.; Nikiforova, S.V.; Avalieri, D.C.; Indo, M.P.; Lockhart, P. The root of flowering plants and total evidence. *Systematic Biology* **2015**, *64*, 879–891. doi:10.1093/sysbio/syv028.
52. Williams, T.A.; Heaps, S.E.; Cherlin, S.; Nye, T.M.W.; Boys, R.J.; Embley, T.M. New substitution models for rooting phylogenetic trees. *Philosophical Transactions of the Royal Society B: Biological Sciences* **2015**, *370*, 20140336.
53. Tarrio, R.; Rodriguez-Trelles, F.; Ayala, F.J. Tree rooting with outgroups when they differ in their nucleotide composition from the ingroup: The Drosophila saltans and willistoni groups, a case study. *Molecular Phylogenetics and Evolution* **2000**, *16*, 344–349. doi:10.1006/mpev.2000.0813.
54. Graham, S.W.; Olmstead, R.G.; Barrett, S.C.H. Rooting phylogenetic trees with distant outgroups: A case study from the commelinoid monocots. *Molecular Biology and Evolution* **2002**, *19*, 1769–1781.
55. Soltis, P.S.; Soltis, D.E. Applying the bootstrap in phylogeny reconstruction. *Statistical Science* **2003**, *18*, 256–267.
56. de la Torre-Barcena, J.E.; Kolokotronis, S.O.; Lee, E.K.; Stevenson, D.W.; Brenner, E.D.; Katari, M.S.; Coruzzi, G.M.; DeSalle, R.; Katari, M.S.; Coruzzi, G.M.; Desalle, R. The impact of outgroup choice and missing data on major seed plant phylogenetics using genome-wide EST data. *Genome biology and evolution* **2009**, *4*, 2–11. doi:10.1371/journal.pone.0005764.
57. Holland, B.R.; Penny, D.; Hendy, M.D. Outgroup misplacement and phylogenetic inaccuracy under a molecular clock - A simulation study. *Systematic Biology* **2003**, *52*, 229–238. doi:10.1080/10635150390192771.
58. Sokoloff, D.D.; Marques, I.; Macfarlane, T.D.; Remizowa, M.V.; Lam, V.K.Y.; Pellicer, J.; Hidalgo, O.; Graham, S.W. Cryptic species in an ancient flowering-plant lineage (Hydatellaceae, Nymphaeales) revealed by molecular and micromorphological data. *Taxon* **2019**, p. in press. doi:10.1002/tax.12026.
59. Roche, D.G.; Kruuk, L.E.B.; Lanfear, R.; Binning, S.A. Public data archiving in ecology and evolution: How well are we doing? *PLOS Biology* **2015**, *13*, e1002295. doi:10.1371/journal.pbio.1002295.
60. DeBiasse, M.B.; Ryan, J.F. Phylotocol: Promoting transparency and overcoming bias in phylogenetics. *Systematic Biology* **2018**, p. in press. doi:10.1093/sysbio/syy090.
61. Gruenstaeudl, M.; Gerschler, N.; Borsch, T. 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* **2018**, *8*, 25. doi:10.3390/life8030025.
62. Whitlock, M.C.; McPeek, M.A.; Rausher, M.D.; Rieseberg, L.; Moore, A.J. Data archiving. *The American Naturalist* **2010**, *175*, 2–3. doi:10.1086/650340.
63. Vines, T.H.; Andrew, R.L.; Bock, D.G.; Franklin, M.T.; Gilbert, K.J.; Kane, N.C.; Moore, J.S.; Moyers, B.T.; Renaut, S.; Rennison, D.J.; Veen, T.; Yeaman, S. Mandated data archiving greatly improves access to research data. *The FASEB Journal* **2013**, *27*. doi:10.1096/fj.12-218164.
64. Vision, T.J. Open data and the social contract of scientific publishing. *BioScience* **2010**, *60*, 330–331. doi:10.1525/bio.2010.60.5.2.
65. 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.; Drummond, A. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **2012**, *28*, 1647–1649. doi:10.1093/bioinformatics/bts199.
66. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* **2013**, *30*, 772–780. doi:10.1093/molbev/mst010.

67. Bi, G.; Mao, Y.; Xing, Q.; Cao, M. HomBlocks: A multiple-alignment construction pipeline for organelle phylogenomics based on locally collinear block searching. *Genomics* **2017**, *110*, 18–22. doi:10.1016/j.ygeno.2017.08.001.
68. Asaf, S.; Khan, A.L.; Khan, M.A.; Shahzad, R.; Lubna.; Kang, S.M.; Al-Harrasi, A.; Al-Rawahi, A.; Lee, I.J. Complete chloroplast genome sequence and comparative analysis of loblolly pine (Pinus taeda L.) with related species. *PLOS ONE* **2018**, *13*, e0192966. doi:10.1371/journal.pone.0192966.
69. Khan, A.; Asaf, S.; Khan, A.L.; Al-Harrasi, A.; Al-Sudairy, O.; Abdulkareem, N.M.; Khan, A.; Shehzad, T.; Alsaady, N.; Al-Lawati, A.; Al-Rawahi, A.; Shinwari, Z.K. First complete chloroplast genomics and comparative phylogenetic analysis of Commiphora gileadensis and C. foliacea: Myrrh producing trees. *PLOS ONE* **2019**, *14*, e0208511. doi:10.1371/journal.pone.0208511.
70. Wakefield, M.J.; Maxwell, P.; Huttley, G.A. Vestige: Maximum likelihood phylogenetic footprinting. *BMC Bioinformatics* **2005**, *6*, 130. doi:10.1186/1471-2105-6-130.
71. Kosakovsky Pond, S.L.; Scheffler, K.; Gravenor, M.B.; Poon, A.F.Y.; Frost, S.D.W. Evolutionary fingerprinting of genes. *Molecular Biology and Evolution* **2010**, *27*, 520–536. doi:10.1093/molbev/msp260.
72. Dubchak, I.; Frazer, K. Multi-species sequence comparison: the next frontier in genome annotation. *Genome Biology* **2003**, *4*, 122.
73. Rosenberg, M.S. Evolutionary distance estimation and fidelity of pair wise sequence alignment. *BMC Bioinformatics* **2005**, *6*, 102. doi:10.1186/1471-2105-6-102.
74. Archie, J.W. Measures of homoplasy. In *Homoplasy: The recurrence of similarity in evolution*; Sanderson, M.J.; Hufford, L., Eds.; Academic Press: Cambridge, MA, 1996; chapter 6, pp. 153–188. doi:10.1016/B978-0-12-618030-5.50008-3.
75. Kluge, A.G.; Farris, J.S. Quantitative phyletics and the evolution of anurans. *Systematic Zoology* **1969**, *18*, 1–32.
76. Farris, J.S. The retention index and the rescaled consistency index. *Cladistics* **1989**, *5*, 417–419.
77. Bofkin, L.; Goldman, N. Variation in evolutionary processes at different codon positions. *Molecular Biology and Evolution* **2007**, pp. 513–521. doi:10.1093/molbev/msl178.
78. Liu, Y.; Cox, C.J.; Wang, W.; Goffinet, B. Mitochondrial phylogenomics of early land plants: Mitigating the effects of saturation, compositional heterogeneity, and codon-usage bias. *Systematic Biology* **2014**, *63*, 862–878. doi:10.1093/sysbio/syu049.
79. Team, R.D.C. *R: A language and environment for statistical computing*; Computing, R Foundation for Statistical: Vienna, Austria, 2013.
80. Paradis, E.; Schliep, K. Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **2018**, *35*, 526–528. doi:10.1093/bioinformatics/bty633.
81. Heibl, C. PHYLOCH: R language tree plotting tools and interfaces to diverse phylogenetic software packages, 2008.
82. Schliep, K.P. Phangorn: Phylogenetic analysis in R. *Bioinformatics* **2011**, *27*, 592–593. doi:10.1093/bioinformatics/btq706.
83. Hurvich, C.M.; Tsai, C.L. Regression and time series model selection in small samples. *Biometrika* **1989**, *76*, 297–307.
84. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **2014**, *30*, 1312–1313. doi:10.1093/bioinformatics/btu033.
85. Stamatakis, A.; Hoover, P.; Rougemont, J. A rapid bootstrap algorithm for the RAxML web servers. *Systematic Biology* **2008**, *57*, 758–771. doi:10.1080/10635150802429642.
86. Ronquist, F.; Huelsenbeck, J.P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **2003**, *19*, 1572–1574.
87. Akaike, H. Likelihood of a model and information criteria. *Journal of Econometrics* **1981**, *16*, 3–14.
88. Fox, J.; Weisberg, S. *An R companion to applied regression*, 2nd ed. ed.; Sage: Thousand Oaks, CA, 2011.
89. Wickham, H. *ggplot2: Elegant graphics for data analysis*, 2nd ed. ed.; Springer: New York, NY, 2016.
90. Kassambara, A. *ggpubr: 'ggplot2' based publication ready plots*, 2018.