Comparative analysis of the complete chloroplast genome of seven *Nymphaea* species

chunqing Sun
Zhen jiang institute of agricultural science in Jiangsu hilly areas

fadi Chen
Nanjing Agricultural University

nianjun Teng
Nanjing Agricultural University

yuemei Yao
Zhenjiang institute of agricultural science in Jiangsu hilly areas

Xi Shan
Zhenjiang institute of agricultural science in Jiangsu hilly areas

zhongliang Dai (✉ daizhongliang2008@126.com)
Zhenjiang institute of agricultural of the science of the Jiang-zhen hilly district

---

Abstract

Background: Although there has been a long history of cultivation and research on *Nymphaea*, the taxonomic relationships and evolutionary relationships among *Nymphaea* species remain controversial. The chloroplast (cp) genome can provide a new method to determine species origin, evolution, and phylogenetic relationships of *Nymphaea*. However, there are few studies on the cp genomes of *Nymphaea*, and the data of their genomes are very scarce. The complete cp genomes of seven *Nymphaea* species were sequenced by high-throughput sequencing technology, and the structural characteristics and phylogenetic relationships of cp genomes were analyzed.

Results: A total of 126−129 genes were annotated in seven *Nymphaea* species, including 81−84 protein coding genes, eight rRNA, and 37 tRNA genes. A comparative cp genomic analysis of seven *Nymphaea* species showed that the cp gene sequence of *Nymphaea* was consistent, with no signs of reverse rearrangement. The codons preferentially ended with A/U. The cp genomes of seven *Nymphaea* species contained 147−168 dispersed repeats with a length of 15−19 bp and 979 simple sequence repeats (SSRs). Using *N. colorata* as the reference sequence, a total of 8,328 single nucleotide polymorphisms (SNPs) and 1,579 insertions/deletions were obtained. The degree of variation of the cp genome of the seven *Nymphaea* species in *rpoA−rpl20*, *rbcL−ndhC*, *ndhD−ndhF*, and *trnN-GUU−ndhA* regions is relatively high; related regions can be used as potential molecular markers for population genetics research. KaKs analysis showed that the *ycf2* gene was positively selected. The results of the phylogenetic analysis showed that the genus *Nymphaea* can be divided into five branches: subgenus *Nymphaea*, subgenus *Hydrocallis*, subgenus *Lotos*, subgenus *Anecphya*, and subgenus *Brachyceras*.

Conclusions: The cp genome of *Nymphaea* is very conservative in structure and composition, but it has rich variation in LSC and SSC regions. The phylogenetic analysis showed that *Nymphaea* could be further divided into five subgenera, and *Euryale* and *Victoria* were most closely related to *Nymphaea*. This study provided insight on the structure and composition of the cp genome of *Nymphaea*. Moreover, our results have provided more sequence information and genome resources for follow-up studies on *Nymphaea*.

Background

Nymphaea, one of the important aquatic flowers, is located at the root of the angiosperm tree [1, 2]. *Nymphaea* consists of approximately 50 species and is distributed on all continents except Antarctica [3, 4]. *Nymphaea* is divided into five subgenera based up on the morphology of flowers, seeds, leaves, and stems, and the geographical distribution, namely, subgenus *Nymphaea*, subgenus *Hydrocallis*, subgenus *Lotos*, subgenus *Anecphya*, and subgenus *Brachyceras* [5]. Subgenus *Anecphya* contains six or seven species only distributed in Australia and New Guinea. There are 10−12 species in subgenus *Brachyceras*, which are widely distributed in the tropics and subtropics. Subgenus *Hydrocallis* is composed of 14 species that are distributed in the neotropics, primarily in South America. Subgenus *Lotos* contains one to three species that are distributed in the tropics of the Old World. Subgenus *Nymphaea* contains six to seven species that are distributed in temperate areas of the Northern Hemisphere. However, there are still

some controversies on the phylogenetic classification of *Nymphaea*. For example, based on the noncoding cp markers, some studies have supported the division of *Nymphaea* genera into three subgenera—Subgenus *Nymphaea* followed by Subgenus *Anecphya-Brachyceras* and another clade comprising Subgenus *Hydrocallis-Lotos* [6, 7]. In addition, relationships among species within *Nymphaea* remain unclear. For example, using the cp *trnT–trnF* sequences to construct phylogenetic trees of Nymphaeaceae, a previous study has supported three subbranches within Subgenus *Nymphaea* [7]. However, a phylogeny was constructed using the internal transcribed spacers (ITS) region of nrDNA and the study supported the division of Subgenus *Nymphaea* into two subbranches [8]. There are different understandings of the evolutionary relationships among *Nymphaea* species; therefore, it is of great significance to study its phylogenetic relationship.

The cp genome is relatively conservative in gene composition and structure compared with the nuclear and mitochondrial genomes [9, 10]. Because of its small size and relatively conservative structure, the cp genome has become an ideal model for evolutionary and comparative genomic research [11], and its structure and sequence information are of great value in revealing the origin, evolution, and relationships among different species. With the rapid development of high-throughput sequencing technology, research on the cp genome has increased [12, 13]. However, there are few studies on the cp genomes of *Nymphaea*, and the data of their genomes are very scarce.

In this study, the cp genomes of seven species of *Nymphaea* were sequenced by high-throughput sequencing technology. The full length of cp genomes was obtained and the composition, structure, and phylogenetic relationships of cp genomes were analyzed in order to enrich the genetic information of *Nymphaea* and provide a theoretical basis for future research on phylogeography, phylogenetics, and conservation of *Nymphaea*.

Results

The basic structural characteristics of the cp genome of seven species of *Nymphaea*

After high-throughput sequencing, raw data was obtained. Low quality reads were removed, leaving the remaining paired-end reads of seven *Nymphaea* species. The paired-end reads ranged from 19,642,319 reads for *N. odorata* to 26,112,273 reads for *N. micrantha*. After performing de novo genome sequencing and assembly, seven complete cp genome sequences for *N. odorata*, *N. rubra*, *N. gigantea*, *N. potamophila*, *N. colorata*, *N. tetragona*, and *N. micrantha* were obtained (Fig 1; Table 1) and submitted to GenBank under the following accession numbers: MT107636, MT107632, MT107637, MT107633, MT107631, MT107634, and MT107635, respectively.

These seven novel cp genome sequences had the classical quadripartite structure that contained one LSC, one SSC, and two IR (IRa and IRb) regions (Fig 1). The cp genomes of seven species of *Nymphaea* had different sizes, among which the largest was *N. gigantea* at 160,179 bp long and the smallest was *N. potamophila* at 159,232 bp (Table 1). The length of the LSC ranged from 89,450–90,266 bp, SSC ranged from 19,340–1,965 bp, and IR ranged from 25,163–25,232 bp (Table 1). Therefore, the variation of the

length of the LSC is greater than that of SSC and IR, and the difference in genome length was mainly due to the varying lengths of the LSC and SSC. The GC content of the seven *Nymphaea* cp genomes was ~39% (Table 1); the GC contents in IR (~43%) of these seven *Nymphaea* species were higher than that of LSC and SSC regions (~37% and 34%).

A total of 126–129 genes were identified in the cp genome of seven *Nymphaea* species, with 126, 127, and 128 in *N. odorata*, *N. rubra*, and *N. tetragona*, respectively, and 129 in the other four species (Table 1). Most of the genes are single-copy genes, but the six protein coding genes (*rps12*, *rpl23*, *rpl2*, *ycf2*, *rps7*, and *ndhB*), seven tRNA genes (*trnI-CAU*, *trnL-CAA*, *trnV-GAC*, *trnI-GAU*, *trnA-UGC*, *trnR-ACG*, and *trnN-GUU*) and four rRNA genes (*rrn16s*, *rrn23s*, *rrn4.5s*, and *rrn5s*) in the IR region contain two copies (Table 2). In the cp genome of the seven *Nymphaea* species, the number of rRNAs and tRNAs was the most conservative, with eight and 37, respectively (Table 1). There were 81–84 protein coding genes (Table 1; Table 2), among which *ycf1* was missing in *N. tetragona*, *ycf4* and atpF were missing in *N. rubra*. *ndhF*, *rpoc2*, and *ycf1* were missing in *N. odorata*.

The IR and SC boundaries of the cp genomes of the seven *Nymphaea* species were compared and analyzed. As shown in Figure 2, the LSC/IRb boundaries of the seven *Nymphaea* species were all within the *rpl2* gene, and the length of the overlapping region between the boundary and the rpl2 gene was 15–39 bp. The IRb/SSC boundaries of the six species were between *trnN* and *ndhF*, and the length of the *ndhF* gene entering SSC was 49–80 bp. However, the SSC region of *N. odorata* lacked an *ndhF* gene and there was a 522 bp gap between the boundary and the *trnN* gene in the IRb region. The SSC/IRa boundaries of five species were all located in *ycf1* gene, the length of the gene in the SSC region was between 5,659–5,719 bp, and was between 155–167 bp in the IRa region. No *ycf1* gene was found in the SSC regions of *N. odorata* or *N. tetragona*, and there were 521 bp and 545 bp gaps between the *trnN* gene in the border and IRa regions, respectively.

Codon preference

The codon preferences of the cp genome of seven *Nymphaea* species were analyzed, and there was no significant difference among the genomes. *N. colorata* was used as an example to determine the codon preference of the cp genome of *Nymphaea*. A total of 25,995 codons were detected in 79 protein coding genes of the *N. colorata* cp genome and their RSCU values were calculated (Additional file 1). Among these codons, 2,636 were used to encode leucine (Leu), accounting for 10.1% of all codons, with UUA being the most commonly used. The content of codons used to code Tryptophan (Trp) was the lowest, 447 in total, accounting for 1.7% of all codons. The RSCU values of 32 codons were greater than 1, of which 28 ended with A/U, indicating that the codon ending preference was A/U. The RSCU of tryptophan (Trp) was equal to 1, indicating that Trp has no codon preference.

Analysis of dispersed repeats and simple repeat sequences (SSRs)

The dispersed repeats of the cp genome include forward, reverse, palindrome, and complementary types. Using Vmatch v2.3.0 software, the dispersed repeat sequences of seven *Nymphaea* species were

identified. For *N. odorata*, *N. rubra*, *N. gigantea*, *N. potamophila*, *N. colorata*, *N. micrantha*, and *N. tetragona*, 163, 147, 151, 158, 153, 155, and 168 repeat sequences, respectively, with a length of 15–19 bp were detected in the cp genome and were distributed in the coding region, the gene interval region, and the gene intron region (Table 3). There was little difference in the total number of dispersed repeats of seven *Nymphaea* species, and the length of the sequences was concentrated between 15–19 bp, accounting for 83.4–88.0% of the total dispersed repeats (Table 3).

A total of 979 SSRs had been identified in the cp genome of seven *Nymphaea* species (Table 3). Among these SSRs, the most abundant was trinucleotide SSR, accounting for 48.2% (472), followed by mononucleotide SSR (44.7%; 438), dinucleotide SSR (3.3%; 32), tetranucleotide SSR (2.9%; 28), pentanucleotide SSR (1.1%; 11), and hexanucleotide SSR (0.1%; 1) (Table 3). Pentanucleotide SSRs were distributed in *N. gigantea* (AAGAA, ATTTA, TATAT, and TTATA), *N. micrantha* (CATAA), *N. odorata* (ATACC), *N. rubra* (AAAGG and TTCCT), and *N. tetragona* (ATACC, TATTA, and TTAGC); a hexanucleotide SSR was only distributed in *N. gigantea* (ATTTAT) (Table 3). Most of the SSRs of the seven species were located in the LSC (56.1–59.7%), followed by the SSC (20.3–21.6%) and the IR (18.8–21.9%).

Collinearity, KaKs, Pi, SNP, and indel analysis

Although the cp genome of plants is conservative, there will be rearrangement over the course of evolution. In this study, we selected the cp genome of *N. colorata* as the reference sequence for comparison with the other six cp genomes of *Nymphaea*. The results showed that there was no reverse rearrangement in the cp genomes of the seven species of *Nymphaea*, indicating a collinear relationship (Fig 3). In addition, the sequence differences of 105–115 kb of the cp genome of the seven species of *Nymphaea* are large (Fig 3); therefore, it can be used to develop molecular markers for future genetic analyses of *Nymphaea*.

The synonymous substitution rate (Ka), missense substitution rate (Ks), and Ka/Ks of seven species of *Nymphaea* were calculated by using the cp genome of *N. colorata* as the reference sequence (Additional file 2). Most of the cp genes have low Ka and Ks, indicating the conservation of the evolutionary process of the cp genome. Ka/Ks < 1 of most genes indicate that they may undergoing purifying selection, i.e., the mutation is harmful and eliminated in the population. Among the genomes, the Ka/KS of *ycf2 in the six comparison groups, N. colorata* vs. *N. gigantea*, *N. colorata* vs. *N. micrantha*, *N. colorata* vs. *N. odorata*, *N. colorata* vs. *N. potamophila*, *N. colorata* vs. *N. rubra*, and *N. colorata* vs. *N. tetragona* were all greater than 1, indicating that *ycf2* was subject to positive selection effect.

The Pi values of eight genes (*petN*, *petG*, *psaI，psbF，psbL，psbM，psbN，*and *rps18*) in the LSC region, all tRNAs, the psaC gene in SSc region, and five genes (*rrn4.5s，rrn5s，rrn16s，rps7，*and *rpl23*) in the IR region were 0 (Additional file 3), indicating that these genes were relatively conserved in the seven *Nymphaea* species. Some genes had relatively high Pi values, for example, *rpl32*, *rps19*, *rpl20*, and *ycf1* had Pi values of 0.012579, 0.011947, 0.009195, and 0.006307, respectively, indicating that the greater the number of polymorphisms in these genes in *Nymphaea*, the more abundant their genetic diversity. These regions of

rpl32–ndhF, *rps19–psbN*, and *atpB–ndhJ* in the cp genome of *Nymphaea* were highly variable, indicating that they may be potential molecular markers (Fig 4).

With *N. colorata* as the reference sequence, 8,328 SNPs and 1,579 insertions/deletions (indels) were obtained by global genome alignment using MAFFT software (Additional file 4; Additional file 5). SNPs were found in both coding and non-coding regions of cp genome. However, there was a relatively large number of SNPs in some highly variable regions located in *rpoA–rpl20*, *rbcL–ndhC*, *ndhD–ndhF*, and *trnN-GUU–ndhA* (Fig 4). These regions may represent differentiation hotspots of cp genome of *Nymphaea*. Moreover, 64.7% of SNPs (5,391) were located in the intergenic region; the remaining SNPs were located in 84 cp genes (Additional file 4). In addition, some genes, such as *matK*, *ndhA*, *ndhF*, *rpoC1*, *rpoc2*, *rpoA*, *rpoB*, *accD*, *atpA*, *ycf1*, and *ycf2*, all showed a high density of SNPs (Fig 4; Additional file 4). We obtained 1,579 indels mutations, 93.3% of which were located in the gene interval and only 6.7% (106) of them were located in the protein coding genes (Additional file 5). These 106 indels were distributed in *matK*, *rps11*, *rpl2*, *rpl22*, *rpl32*, *rpoC2*, *petB*, *ndhF*, *ndhJ*, *accD*, *atpE*, *ycf1*, *ycf2*, and *ycf4* genes.

Phylogenetic analysis

Nymphaeaceae was divided into five main branches, *Brasenia*, *Nelumbo*, *Nuphar*, *Nymphaea*, and *Cabomba*, but *Euryale* and *Victoria* were closely related to *Nymphaea* (Fig 5). *Nymphaea* can be further divided into five subbranches, corresponding to the five *Nymphaea* subgenera (Fig 5). The first branch consists of *N. gigantea* from the subgenus *Anecphya*; the second subfamily consists of *N. colorata*, *N. micrantha*, *N. capensis*, and *N. ampla* which compose subgenus *Brachyceras*. The third branch is the subgenus *Nymphaea*, which is composed of *N. mexicana*, *N. odorata*, *N. alba var. rubra*, *N. alba*, and *N. tetragona*; *N. potamophila* and *N. jamesoniana* compose the fourth branch, subgenus *Hydrocallis*. The fifth branch is the subgenus *Lotos*, which contains of *N. lotus* and *N. rubra*.

Discussion

The structure and composition of cp genome of *Nymphaea*

The cp genome of *Nymphaea* is composed of LSC, SSC, IRa, and IRb regions, and the overall length of the cp genome within middle and upper ranges of angiosperms (159–160 kb) [14]. Although the structure and composition of the cp genome of seven *Nymphaea* are very similar, a comparative analysis shows that there are also many variations in the microstructure. The SSC and LSC regions contain the most variation, while multiple copies of the two IR regions make the IR region less prone to mutation [15]. In this study, *N. gigantea* and the other six species had a large difference in cp genome length, while *N. odorata* and *N. tetragona*, as well as *N. colorata* and *N. micrantha* had similar cp genome length, which revealed that *N. gigantea* was distant from the other six species, while *N. odorata* and *N. tetragona*, *N. colorata* and *N. micrantha* were closely related. The difference in the cp genome size was mainly caused by the contraction and expansion of the IR [14]. However, the difference of the cp genome size of *Nymphaea* was primarily related to the length variation of LSC and SSC. Gene loss in the cp genome was

common, for example, gene *ycf1* and *ycf2* were completely lost in *Gramineae* [16], and gene *infA* was lost in many angiosperm cp genomes [17, 18]. In this study, the cp genome of seven *Nymphaea* species also had gene loss and the *ycf1* gene disappeared completely in *N. odorata* and *N. tetragona*. It is unknown whether *ycf1* was transferred to the nuclear or mitochondrial genome. The GC content of the cp genomes of seven *Nymphaea* species showed differences in four regions, with the highest content in the IR region (43.35–43.44%), while the GC content in the LSC and SSC regions was lower, which is due to the existence of rRNA and tRNA in IR region. When the GC content in genome is greater, it has a greater the density of DNA bases, making the DNA sequence more stable and difficult to mutate [14].

The IR region is the most conservative region in the cp genome, but it also shrinks and expands during evolution [19]. Gene *trnH-GUG* and *ndhF* often appeared at the boundary of IR and SC region of cp gene, which were considered to be footprints of cp genome evolution [20]. We analyzed the contraction and expansion of IR region and found that the gene distribution at the boundary of the four regions of the cp genome had the same rule, but differences were observed in the microstructure, especially the locations of *rps19*, *ndhF*, *ycf1*, *and trnH-GUG*. According to the different locations of these four genes, *N. tetragona* and *N. odorata* can be identified from the other *Nymphaea* species. The boundary of SSC and IRa region is located in gene *ycf1*. The expansion of the IR region into *ycf1* had also been described in *Cardiocrinum* [21] and *Amana* [22]. Therefore, *rps19*, *ndhF*, *ycf1*, and *trnH-GUG* can be considered part of the evolutionary footprint of the cp genome of *Nymphaea.*

Codon preference

In this study, 67 codons were used in seven *Nymphaea* species, with 32 preferred codons (RSCU>1). The most commonly used codon in the cp genome is the codon with a higher AU content [23], and the preferred codon tends to end with A/U [24]. In this study, only UUG and AUG codons of the 32 preferred codons were G-terminal and the other preferred codons were A/U-terminal. Leucine (Leu) was the most encoded amino acid, and its synonymous codon preference was UUA>UUG>CUU>CUA>CUG>CUC, which is inconsistent previous studies conducted in other families [16, 25]. The codon usage is related to gene expression level, base composition, asymmetric mutation of DNA strands, and selection [26], but the mechanism of codon usage requires further study. The results of this study provide insight into the evolution of the cp genome of *Nymphaea.*

Repeat sequence analysis

In this study, the length of the genome repeat sequence of seven *Nymphaea* species was between 15–19 bp and no large segment repeat sequence (>100 bp) reported in other species was found [27, 28]. Large fragment repeats are related to gene rearrangement [29]. For example, *Erodium*, *Pelargonium*, and *Geranium* (Geraniaceae) have gene rearrangements that all contain large fragment repeats (>100 bp) [9]. The results of the collinearity analysis of seven *Nymphaea* species showed that all the genes are collinear and no rearrangement occurs, which is consistent with the result that there is no large segment repeat sequence in cp genome.

SSR markers are highly polymorphic and are therefore used as molecular markers in population genetics and studies of evolution. Most SSRs in the cp genomes of seven *Nymphaea* species are trinucleotide SSRs (48.2%), which is inconsistent with the main SSR types of other plants [30, 31]. For example, most SSRs in the cp genomes of *Primula* are single nucleotide SSRs [25]. In addition, in the gene coding region, the number of SSRs in *ycf1* is the largest, which is consistent with the research conducted in *Primula* [25] and *Cardiocrinum* [21]. The SSRs in the cp genome reported in this study can be used as potential molecular markers in future studies on *Nymphaea*.

Nucleotide polymorphism, KaKs, SNP and indel analysis

The Pi values of the LSC and SSC regions were significantly higher than that of the IR region, which was related to the conservation of the IR region. Three gene spacer regions, *rpl32–ndhF*, *rps19–psbN*, and *atpB–ndhJ*, as well as four protein coding regions, *rpl32*, *rps19*, *rpl20*, and *ycf1*, have been reported in other plants as highly variable regions, and they are also considered potential molecular markers [27, 32-33]. These regions may undergo faster nucleotide substitution rates during the course of evolution, which is of great significance to phylogenetic analysis and identification of *Nymphaea*.

Our result showed that *ycf2 in the* cp genome of *Nymphaea* was under positive selection. *ycf2* was located in the reverse repeat region of cp genome, which is an important enzyme coding gene in the cp [34]. The evolutionary rate of *ycf2* is relatively fast, its gene function is controversial, and it has not been classified into a main functional gene type (genetic system gene and photosynthetic system gene) [35]. However, because of its cp ATPase coding function [34], it is believed to have a regulatory role in the development of fruit and flower organs [35], which is considered one of the important genes in the cp genome. Sites undergoing positive selection were detected in gymnosperms and other angiosperms, indicating that the evolution of *ycf2* is widespread in plants [36]. The seven *Nymphaea* species in this study are distributed on different continents and grow under quite different environmental conditions. Therefore, the positive selection of *ycf2 in Nymphaea* may be related to the difference of their optimal growth environment and the different demand for light intensity. The discovery of positive selection on *ycf2* confirmed that there was evolution at the molecular level in the process of environmental adaptation for *Nymphaea*.

Large indels are uncommon among the cp genomes of the same genus [15]. Similarly, for the cp genome of *Nymphaea*, most of indels exist in the noncoding region, and the largest indel is only 28 bp long. Compared with the small number of indels, we found 8,328 SNPs in the cp genome of *Nymphaea*. These SNPs and indel regions have a rapid rate of evolution and the degree of variation was the greatest in the *rpoA–rpl20*, *rbcL–ndhC*, *ndhD–ndhF*, and *trnN-GUU–ndhA* intergenic regions. These data indicate that these regions may be suitable DNA barcodes for *Nymphaea*.

Phylogenetic analysis

In this study, a phylogenetic tree were constructed with 21 species of seven genera in Nymphaeaceae. *Cabomba* was found to be at the base of the phylogenetic tree. *Euryale*, *Victoria*, and *Nymphaea*

subgenus *Hydrocallis* and subgenus *Lotos* formed a clade, indicating that the three genera are closely related. Using *rbcL*, *matK*, and 18S *rRNA* sequences to construct phylogenetic trees of Nymphaeaceae, a previous study has supported the establishment of Cabombaceae by separating *Cabomba* and *Brasenia* from Nymphaeaceae [37]. However, *Cabomba* and *Brasenia* are not sister to each other in this study, indicating that it is not appropriate to combine them to form Cabombaceae. At present, most researchers agree that Nymphaeaceae is composed of eight genera, *Nuphar*, *Barclaya*, *Ondinea*, *Nymphaea*, *Euryale*, *Brasenia*, *Victoria*, and *Cabomba* [37-39]. Our results are consistent with the previous classification of Nymphaeaceae.

Phylogenetic trees were constructed for 15 *Nymphaea* species and the results showed that *Nymphaea* can be divided into five subgenera, which is consistent with the previous classification [8]. However, the relationship between species in the genus *Nymphaea* is still unclear and needs further study. In this study, subgenus *Anecphya*, subgenus *Hydrocallis*, and subgenus *Lotos* form separate branches indicating support for these subgenera, but subgenus *Nymphaea* and subgenus *Brachyceras* showed further divisions. Subgenus *Brachyceras* can be divided into two branches, one containing *N. capensis* and *N. colorata*, the other composed of *N. ampla* and *N. micrantha.* Subgenus *Nymphaea* can also be divided into two branches, one containing *N. odorata* and *N. mexicana*, and the other containing *N. alba* and *N. tetragona*. The division of subgenus *Nymphaea* and subgenus *Brachyceras* may be caused by different growing environments. Our results are consistent with the results presented by Dkhar et al. [8]. The genetic relationship between subgenus *Nymphaea* and subgenus *Brachyceras*, subgenus *Brachyceras* and subgenus *Anecphya*, subgenus *Hydrocallis* and subgenus *Lotos* is relatively close, which is also supported by their cross compatibility [40-42].

Conclusions

The complete cp genomes of seven Nymphaea species were sequenced and successfully constructed. The cp genome of Nymphaea is very conservative in structure and composition, but it has rich variation in LSC and SSC regions, which can be used for germplasm identification, phylogenetic analysis, and evolution research. The phylogenetic analysis showed that Nymphaea could be further divided into five subgenera, and Euryale and Victoria were most closely related to Nymphaea. This study not only enriched the genetic information on Nymphaea, but also laid a theoretical foundation for the evaluation of water lily germplasm resources, molecular breeding, development of SSR molecular markers, and research into genetic diversity. Moreover, this study provided support for the resolution of the phylogenetic relations among Nymphaea species.

Methods

Plant materials

Seven *Nymphaea* species, *N. odorata*, *N. rubra*, *N. gigantea*, *N. potamophila* *N. tetragona*, *N. colorata*, and *N. micrantha*, were included in this study. Herbarium specimens are all preserved at the Zhenjiang

Agricultural Science and Technology Innovation Center and live plants were introduced to the Shanghai Chenshan Botanical Garden. These plants were identified by Nianjun Teng and Yingchun Xu of Nanjing Agricultural University according to their key morphological characteristics provided in Huang et al. [5], and the voucher ID were listed in Table 1. In August 2019, fresh leaves were collected, wrapped with tinfoil, frozen in liquid nitrogen, and stored at -80°C.

Extraction and sequencing of whole genome DNA

Whole genome DNA was extracted with the plant genomic DNA Extraction Kit (TIANGEN Beijing China). After testing the genomic DNA of the sample, DNA was fragmented by sonication, and then purified by fragment purification, terminal repair, 3'-terminal plus A, linked sequencing adapter, and agarose gel electrophoresis for fragment size selection, and PCR amplified into a sequencing library. After passing quality inspection, the constructed library was sequenced with an Illumina Novaseq platform. After sequencing, raw data was filtered out to remove the joint sequence and low-quality reads to obtain high-quality clean data.

Cp genome assembly

First, we used Bowtie 2 V2.2.4 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) software to compare the cp genome database built by Nanjing Jisi Huiyuan Biotechnology Co., Ltd. and selected the matching clean reads for subsequent assembly. We then assembled the filtered clean reads with SPAdes V3.10.1 (http://cab.spbu.ru/software/spades/) software. If the cyclic gene sequence could be directly obtained, the sequence was genome corrected, and then the corrected genome was rearranged to obtain the complete cyclic cp genome sequence. If the complete circular genome was not directly obtained, we used SSPACEV2.0 (http://www.baseclear.com/services/bioinformatics/basetools/sspace-standard/) software to connect the contig sequences to obtain the scaffold sequence; Gapfiller V2.1.1 (http://www.sourceforge.net/projects/gapfiller/) software was used to supplement gaps If gaps still exist, primers were designed, sequenced by PCR, reassembled until the complete pseudo genome sequence was obtained, and then the sequences were matched to the pseudo genome for genome correction. Finally, the corrected sequence was rearranged according to the cp structure to obtain the complete circular cp genome sequence.

Structural annotation of the cp genome

We used Blast v2.2.25 (http://blast.ncbi.nlm.nih.gov/Blast.cgi) software to compare the coding DNA sequence (CDS) of the cp genome in the NCBI database (http:// www.ncbi.nlm.nih.gov/genomes). The final cp genome gene annotation was obtained after manual correction. rRNA annotation information for cp genome was obtained using hmmer v3.1b2 (http://www.hmmer.org/) software to compare the rRNA sequence of cp genome in NCBI database. We used Aragorn v1.2.38 (http://130.235.244.92/aragorn/) software to predict the tRNA of the cp genome sequence and obtain the tRNA annotation information of the cp genome. OGDRAW (http://chlorobox.mpimp-golm.mpg.de/OGDraw.html) software was used to visualize the annotation results and draw a physical map of the cp genome.

Analysis of the cp genome

IR/SC (Inverted repeat/Single-copy region) boundary comparison and codon preference analysis of cp genomes

There are four boundaries between IR (Inverted repeat), LSC (Large single-copy region), and SSC (Small single-copy region). In the process of genome evolution, the IR boundary will expand and contract, pushing some genes into the IR region or a single copy region. We used Adobe Illustrator CS5 mapping software to map the annotation information of the cp genome of seven species of *Nymphaea* to the map of the simplified structure of the cp genome, which was used to compare the boundaries of four regions of the cp genome of *Nymphaea*, in order to show the extent of the shrinking and expanding of IR regions. The number of codons encoded by all protein coding genes and the relative synonymous codon usage (RSCU) in the cp genome of seven species of *Nymphaea* were calculated using Codonw software (http://codonw.sourceforge.net/). RSCU value > 1 indicates that the codon is a preferred codon and is frequently used. RSCU = 1 indicates that the codon has no preference. The value of RSCU < 1 indicates that the frequency of using the codon is low.

Repeat sequence and SSR Analysis

We used Vmatch v2.3.0 (http://www.vmatch.de/) software to identify dispersed repeats. The SSR analysis, including single nucleotide SSRs, dinucleotide SSRs, trinucleotide SSRs, tetranucleotide SSRs, pentanucleotide SSRs, and hexanucleotide SSRs, was conducted using MISA V1.0 (MIcroSAtellite identification tool; http://pgrc.ipkgatersleben.de/misa/misa.html) software. The minimum repetition times of these SSRs were set to eight, five, three, three, three, and three, respectively.

KaKs, nucleotide polymorphism, and global comparison analysis

The gene sequence was compared using MAFFT V7.310 (https://mafft.cbrc.jp/alignment/software/) software, and the values of Ka and Ks were calculated by KaKs-Calculator V2.0 (https://sourceforge.net/projects/kakscalculator2/) software. Ka/Ks > 1 indicates positive selection; Ka/Ks < 1 indicates purifying selection. The CDS sequences of the same gene in seven species of *Nymphaea* were compared globally using the MAFFT software, and the nucleotide polymorphism (Pi) of each gene was calculated using VCFtools. The cp genomes of seven species of *Nymphaea* were compared globally using MAFFT software, and then SNPs and indel were identified.

Cp genome collinearity and phylogenetic tree analysis

Using the cp genome of *N. colorata* as the reference sequence, the cp genomes of seven novel sequences from this study were compared using Mauve software to detect whether there was rearrangement in the cp genomes among the included *Nymphaea* species. Seven cp genes of the genus *Nymphaea* in this study and 14 Nymphaeaceae species published in NCBI database were selected for phylogenetic analysis. We carried out multiple sequence alignment using MAFFT software. The evolutionary tree was

built using RAxMLv8.2.10 (https://cme.h-its.org/exelixis/software.html) software, and the GTR model and hill clipping algorithm were selected.

Abbreviations

Cp: chloroplast; SSR: simple sequence repeat; SNP: Single nucleotide polymorphism; CDS: Coding DNA sequence; IR: Inverted repeat; SC: Single-copy region; LSC: Large single-copy region; SSC: Small single-copy region; RSCU: Relative synonymous codon usage; Pi: Nucleotide polymorphism; Ka: synonymous substitution rate; Ks: missense substitution rate; ITS: internal transcribed spacers.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The seven cp genomes sequences we obtained from this study were archived in NCBI. The accession numbers are MT107636, MT107632, MT107637, MT107633, MT107631, MT107634, and MT107635.

Competing interests

The authors declare that they have no competing interests.

Funding

Authors' contributions

CQS, FDC, NJT and ZLD designed the experiments. CQS and XS performed the experiments. CQS, NJT and ZLD analyzed the data. CQS wrote the manuscript. All authors have read and approved the final manuscript.

Acknowledgements

Author details

1 Zhenjiang Institute of Agricultural Science in Jiangsu Hilly Areas, Jurong, 242400, China.

2 College of Horticulture, Nanjing Agricultural University, Nanjing 210095, China.

References

1. 1. Qiu YL, Lee J, Bernasconi-Quadroni F, Soltis DE, Chase M W. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. Nature. 1999; 402(6760): 404–7.

2. Zanis MJ, Soltis DE, Soltis PS, Mathews S, Donoghue MJ. The root of the angiosperms revisited. Proceedings of the National Academy of Sciences of the United States of America. 2002; 99(10): 6848–53.

3. 3. Sun CQ, Ma ZH, Sun GS, Dai ZL, Teng NJ, Pan YP. Cellular mechanisms of reproductive barriers in some crosses of water lily (*Nymphaea*) cultivars. Hort Science. 2015; 50:30–5.

4. 4. Sun CQ, Ma ZH, Zhang ZC, Sun GS, Dai ZL. Factors influencing cross barriers in interspecific hybridizations of water lily. Journal of the American Society for Horticultural Science. 2018; 143(2):1–6.

5. 5. Huang GZ, Deng HQ, Li Z, Li G. Water lily. Beijing: China Forestry Press; 2009.

6. 6. Löhne C, Borsch T, Wiersema JH. Phylogenetic analysis of Nymphaeales using fast-envolving and noncoding chloroplast Bot J Linn Soc. 2007; 154:141–63.

7. Borsch T, Hilu KW, Wiersema JH, Löhne C, Barthlott W, Wilde V. Phylogeny of *Nymphaea* (Nymphaeaceae): evidence from substitutions and microstructural changes in the chloroplast trnT-trnF region. Int J Plant Sci. 2007; 168:639–71.

8. Dkhar J, Kumaria S, Rao SR, Tandon P. Sequence characteristics and phylogenetic implications of the nrDNA internal transcribed spacers (ITS) in the genus *Nymphaea* with focus on some indian representatives. Plant Systematics and Evolution. 2012; 298(1): 93–108.

9. Guisinger MM, Kuehl JV, Boore JL, Jansen RK. Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. Molecular biology and evolution. 2011; 28(1):583–600.

10. Blazier JC, Ruhlman TA, Weng ML, Rehman SK, Sabir JSM, Jansen RK. Divergence of RNA polymerase α subunits in angiosperm plastid genomes is mediated by genomic rearrangement. Sci Rep. 2016; 6: 24595.

11. Dong WP, Xu C, Cheng T, Lin K, Zhou SL. Sequencing angiosperm plastid genomes made easy: a complete set of universal primers and a case study on the phylogeny of Saxifragales. Genome Biol Evol. 2013; 5: 989–97.

12. Lee SR, Kim K, Lee BY, Lim CE. Complete chloroplast genomes of all six *Hosta* species occurring in Korea: molecular structures, comparative, and phylogenetic analyses. BMC Genomics. 2019; 20:833.

13. Duan H, Guo JB, Xuan L, Wang ZY, Li MZ, Yin YL, Yang Y. Comparative chloroplast genomics of the genus *Taxodium.* BMC Genomics. 2020; 21:114.

14. Terakami S, Matsumura Y, Kurita K, Kanamori H, Katayose Y, Yamamoto T, Katayama H. Complete sequence of the chloroplast genome from pear (*Pyrus pyrifolia*): genome structure and comparative analysis. Tree Genet Genom. 2012; 8:841−54.

15. Jansen RK, Christopher S, Seung-Bum L, Hansen AK, Henry D. Complete plastid genome sequences of three rosids (*Castanea*, *Prunus*, *Theobroma*): evidence for at least two independent transfers of *rpl22* to the nucleus. Molecular Biology and Evolution. 2011; 28(1): 835−47.

16. Guisinger MM, Chumiley TW, Kuehl JV, Boore JL, Jansen RK. Implications of the plastid genome sequence of Typha (Typhaceae, Poales) for understanding genome evolution in Poaceae. Journal of Molecular Evolution. 2010; 70(2):149−66.

17. Gichira AW, Li ZZ, Saina JK. The complete chloroplast genome sequence of an endemic monotypic genus *Hagenia* (Rosaceae): structural comparative analysis, gene content and microsatellite detection. Peer j. 2017; 5(6):e2846.

18. Liu LX, Li R, Worth JRP, Xian L, Pan L, Gameron KM, Fu CX. The complete chloroplast genome of Chinese Bayberry (*Morella rubra*, Myricaceae): implications for understanding the evolution of Fagales. Frontiers in Plant Science. 2017; 8:968.

19. Kim KJ, Lee HL. Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. DNA Research. 2004; 11(4):247−61.

20. Lin CP, Wu CS, Huang YY, Chaw SM. The complete chloroplast genome of *Ginkgo biloba* reveals the mechanism of inverted repeat contraction. Genome biology and evolution. 2012; 4(3):374−81.

21. Lu RS, Li P, Qiu YX. The complete chloroplast genomes of three *Cardiocrinum* (Liliaceae) species: comparative genomic and phylogenetic analyses. Frontiers in Plant Science. 2017; 7:2054.

22. Li P, Lu RS, Xu WQ, Ohi-Toma T, Cai MQ, Qiu YX, Cameron KM, Fu CX. Comparative genomics and phylogenomics of east Asian Tulips (*Amana*, Liliaceae). Frontiers in Plant Science. 2017; 8: 451.

23. Shimada H, Sugiura M. Fine structural features of the chloroplast genome: comparison of the sequenced chloroplast genomes. Nucleic Acids Res. 1991; 19: 983−95.

24. Zhou M, Long W, Li X. Patterns of synonymous codon usage bias in chloroplast genomes of plants. Forest Ecosystems. 2008; 4:235−42.

25. Ren T, Yang YC, Zhou T, Liu ZL. Comparative plastid genomes of Primula species: sequence divergence and phylogenetic relationships. International Journal of Molecular Sciences. 2018; 19(4):1050.

26. Morton BR. Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages. Journal of Molecular Evolution. 1998; 46(4):449−59.

27. Yang Y, Zhou T, Duan D, Feng L, Zhao G. Comparative analysis of the complete chloroplast genomes of five Quercus species. Frontiers in plant science. 2016; 7: 959.

28. Hu Y, Woeste KE, Zhao P. Completion of the chloroplast genomes of five Chinese *Juglans* and their contribution to chloroplast phylogeny. Frontiers in Plant Science. 2016; 7: 1955.

29. Weng ML, Blazier JC, Govindu M, Jansen RK. Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats, and nucleotide substitution rates. Molecular Biology and Evolution. 2013; 31(3):645–59.

30. Kuang DY, Wu H, Wang YL, Gao LM, Zhang SZ, Lu L. Complete chloroplast genome sequence of *Magnolia kwangsiensis* (Magnoliaceae): implication for DNA barcoding and population genetics. Genome. 2011; 54: 663–73.

31. Martin G, Baurens FC, Cardi C, Aury JM, D'Hont A. The complete chloroplast genome of banana (*Musa acuminata*, Zingiberales): insight into plastid monocotyledon evolution. PloS One. 2013; 8:e67350.

32. Shaw J, Shafer HL, Leonard OR, Kovach MJ, Schorr M, Morris AB. Chloroplast DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in angiosperms: the tortoise and the hare IV. American journal of Botany. 2014; 101(11):1987–2004.

33. Niu Z T, Xue QY, Zhu SY, Sun JJ, Liu W, Ding XY. The complete plastome sequences of four orchid species: insights into the evolution of the orchidaceae and the utility of plastomic mutational hotspots. Frontiers in plant science. 2017; 8:715.

34. Wolfe KH. Similarity between putative ATP-binding sites in land plant plastid ORF2280 proteins and the FtsH/CDC48 family of ATPases. Current Genetics. 1994; 25(4):379–83.

35. Drescher A, Ruf S, Calsa TJ, Carrer H, Bock R. The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. The Plant Joural. 2000; 22(2):97–104.

36. Wang B, Gao L, Su YJ, Wang T. Adaptive evolutionary analysis of chloroplast genes in euphyllophytes based on complete chloroplast genome sequences. Acta Scientiarum Naturalium Universitatis Sunyatseni. 2012; 51(3):108–13.

37. Les D H, Schneider E L, Padgett D J. Phylogeny, classification and floral evolution of water lilies (Nymphaeales): A synthesis of non-molecular, rbcL, matK and 18S rDNA data. Systematic Botany. 1999; 24:28–46.

38. Tamura M. Relationship of Barclaya and classification of Nymphaeales. Acta phytotaxonomica et geobotanica. 1982; 33:336–45.

39. Moseley MF, Schneider EL, Williamson PS. Phylogenetic interpretations from selected floral vasculature characters in the Nymphaeaceae sensu lato. Aquatic Botany. 1993; 44:325–42.

40. Songpanic P, Hongtrakul V. Intersubgeneric cross in *Nymphaea* spp. L. to develop a blue hardy waterlily. Scientia Horticulturae. 2010; 124:475–81.

41. Li SJ, Yu Q, Chen C, Zhang Y, Wu YP, Yu G. Breeding progress of waterlilies in China. Journal of plant genetic resources. 2019; 20(4): 829–35.

42. Sun CQ, Chen FD, Teng NJ, Yao YM, Shan X, Dai ZL. Transcriptomic and proteomic analysis reveals mechanisms of low pollen-pistil compatibility during water lily cross breeding. BMC Plant Biology. 2019; 19:542.

Tables

Table 1 Summary statistics for the assembly of seven *Nymphaea* species chloroplast genomes

Genome features	*N. odorata*	*N.tetragona*	*N. rubra*	*N. gigantea*	*N. potamophila*	*N. colorata*	*N. micrantha*
Voucher ID	ZJWL-031	ZJWL-036	ZJWL-008	ZJWL-014	ZJWL-021	ZJWL-041	ZJWL-043
NCBI accession NO.	MT107636	MT107634	MT107632	MT107637	MT107633	MT107631	MT107635
Genome size（bp）	159968	159956	159322	160179	159232	159842	159900
LSC size（bp）	90025	89837	89564	90266	89450	89969	89950
SSC size（bp）	19533	19655	19340	19541	19456	19547	19576
IR size （bp）	25205	25232	25209	25186	25163	25163	25187
Number of genes	126	128	127	129	129	129	129
Protein genes [unique]	81(75)	83（77）	82(76)	84(78)	84(78)	84(78)	84(78)
tRNA genes [unique]	37(30)	37（30）	37(30)	37(30)	37(30)	37(30)	37(30)
rRNA genes [unique]	8(4)	8（4）	8(4)	8(4)	8(4)	8(4)	8(4)
Duplicated genes in IR	32	32	32	32	32	32	32
GC content (%)	39.08	39.14	39.09	39.11	39.13	39.14	39.16
GC content in LSC (%)	37.76	37.82	37.73	37.70	37.79	37.77	37.80
GC content in SSC (%)	34.19	34.31	34.20	34.44	34.28	34.42	34.38
GC content in IR (%)	43.35	43.37	43.37	43.44	43.39	43.43	43.44

Table 2 List of genes in *N. gigantea, N. potamophila，N. colorata* and *N. micrantha* chloroplast genomes.

Gene classification	Gene group	Gene name
Self-replication	Ribosomal RNAs	*rrn16s(×2), rrn23s(×2), rrn4.5s(×2), rrn5s(×2),*
	Transfer RNAs	*trnH-GUG, trnK-UUU, trnQ-UUG, trnS-GCU, trnG-UCC, trnR-UCU, trnC-GCA, trnD-GUC, trnY-GUA, trnE-UUC, trnT-GGU, trnS-UGA, trnG-GCC, trnfM-CAU, trnS-GGA, trnT-UGU, trnL-UAA, trnF-GAA, trnV-UAC, trnM-CAU, trnW-CCA, trnP-UGG, trnI-CAU(×2), trnL-CAA(×2), trnV-GAC(×2), trnI-GAU(×2), trnA-UGC(×2), trnR-ACG(×2), trnN-GUU(×2), trnL-UAG*
	Ribosomal protein (SSU)	*rps16, rps2, rps14, rps4, rps18, rps12(×2), rps3, rps11, rps8, rps19, rps7(×2), rps15,*
	Ribosomal protein (LSU)	*rpl33, rpl20, rpl36, rpl14, rpl16, rpl2(×2), rpl23(×2), rpl32、rpl22*
	RNA polymerase	*rpoC1, rpoC2, rpoB, rpoA*
Gene for photosynthesis	Photosystem 1	*psaA, psaB, psaC, psaI, psaJ*
	Photosystem 2	*psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ*
	Cytochrome b/f complex	*petA, petB, petD, petG, petL, petN*
	ATP synthase	*atpA, atpB, atpE, atpF, atpH, atpI*
	ATP-dependent protease subunit p gene	*clpP*
	Large subunit of rubisco	*rbcL*
	NADH dehydrogenase	*ndhA, ndhB(×2), ndhC, ndhD, ndhE, ndhF, ndhG, ndhI, ndhJ, ndhK*
Other genes	Maturase	*matK*
	Envelop membrane protein	*cemA*
	Subunit of acetyl-CoA-carboxylase	*accD*
	c-type cytochrome synthesis ccsA gene	*ccsA*
	Translation initiation factor IF-1	*infA*
Gene for unknown function	Hypothetical chloroplast reading frames	ycf1, ycf2*(×2)*, ycf3, ycf4

Table 3 Summary of dispersed repeats and SSRs in seven *Nymphaea* species

| Species		N. tetragona	N. odorata	N. rubra	N. gigantea	N. potamophila	N. colorata	N. micrantha
Dispersed repeats	15-19 bp	168	163	147	151	158	153	155
	20-24 bp	24	11	14	27	16	18	20
	25-29 bp	7	7	5	1	7	3	7
	30-60 bp	79	81	76	96	80	6	87
Mononucleotides SSR		61	72	63	56	59	61	63
Dinucleotides SSR		4	4	6	5	5	4	4
Trinucleotides SSR		70	71	69	65	71	64	62
Tetranucleotides SSR		2	2	7	5	4	4	4
Pentanucleotides SSR		3	1	2	4	0	0	1
Hexanucleotides SSR		0	0	0	1	0	0	0
SSR in LSC		81	89	83	77	78	78	77
SSR in SSC		30	32	31	29	31	27	29
SSR in IR		28	28	32	28	30	28	28

Figures
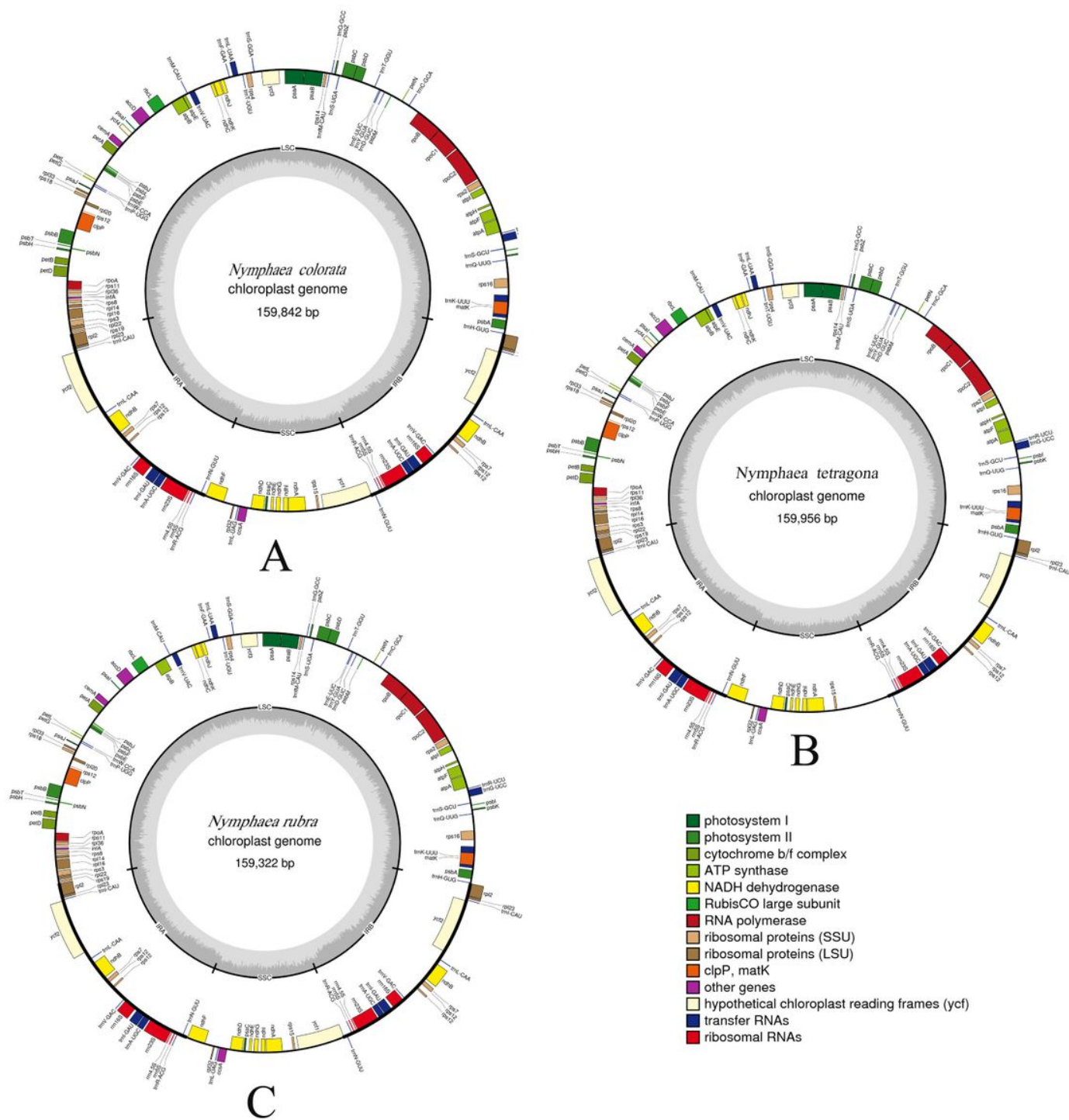
Figure 1

Seven novel cp genome sequences had the classical quadripartite structure that contained one LSC, one SSC, and two IR (IRa and IRb) regions
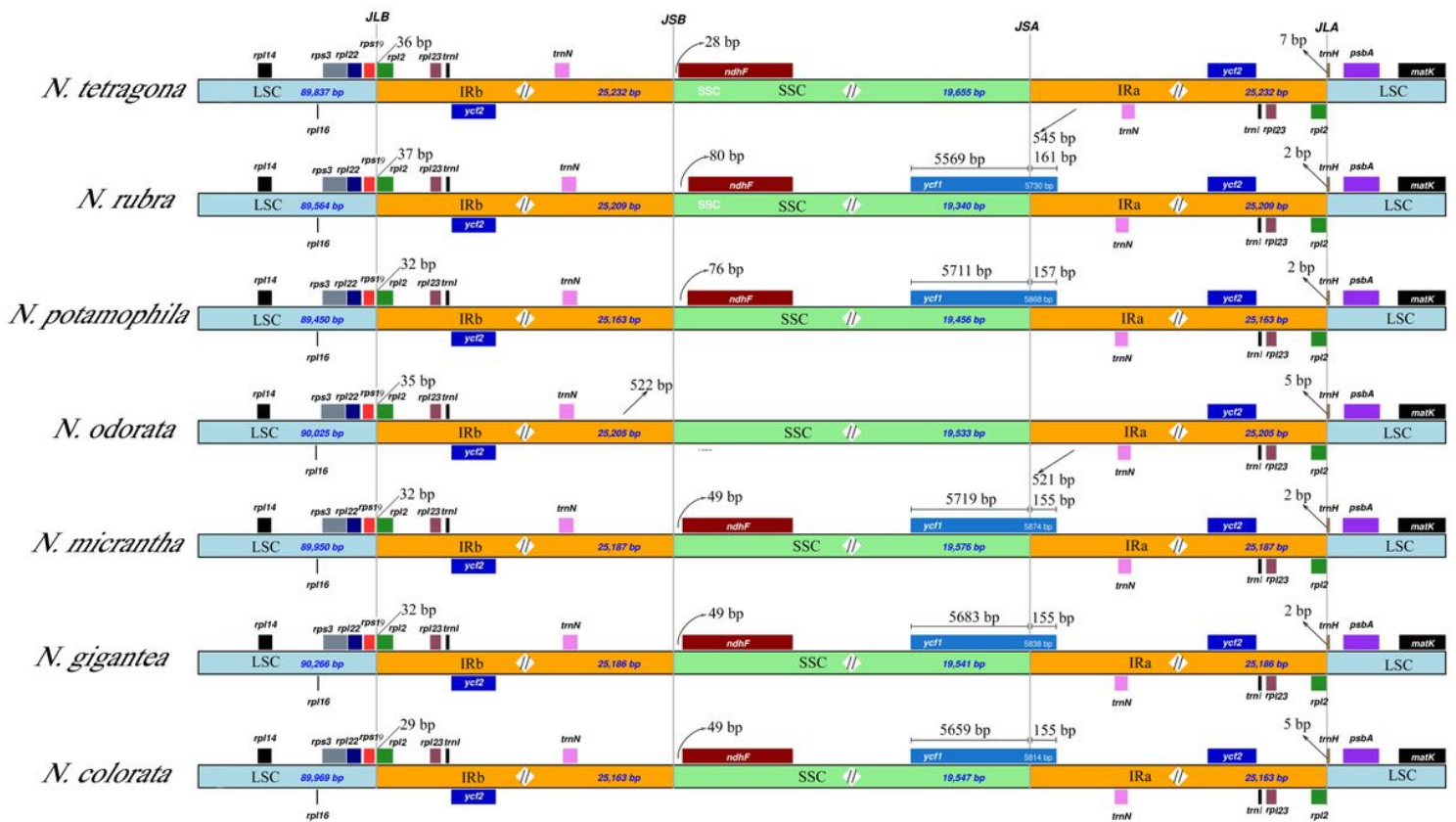
Figure 2

The LSC/IRb boundaries of the seven Nymphaea species were all within the rpl2 gene, and the length of the overlapping region between the boundary and the rpl2 gene was 15−39 bp.

Figure 3

The sequence differences of 105−115 kb of the cp genome of the seven species of Nymphaea

Figure 4

A relatively large number of SNPs in some highly variable regions located in rpoA–rpl20, rbcL–ndhC, ndhD–ndhF, and trnN-GUU–ndhA
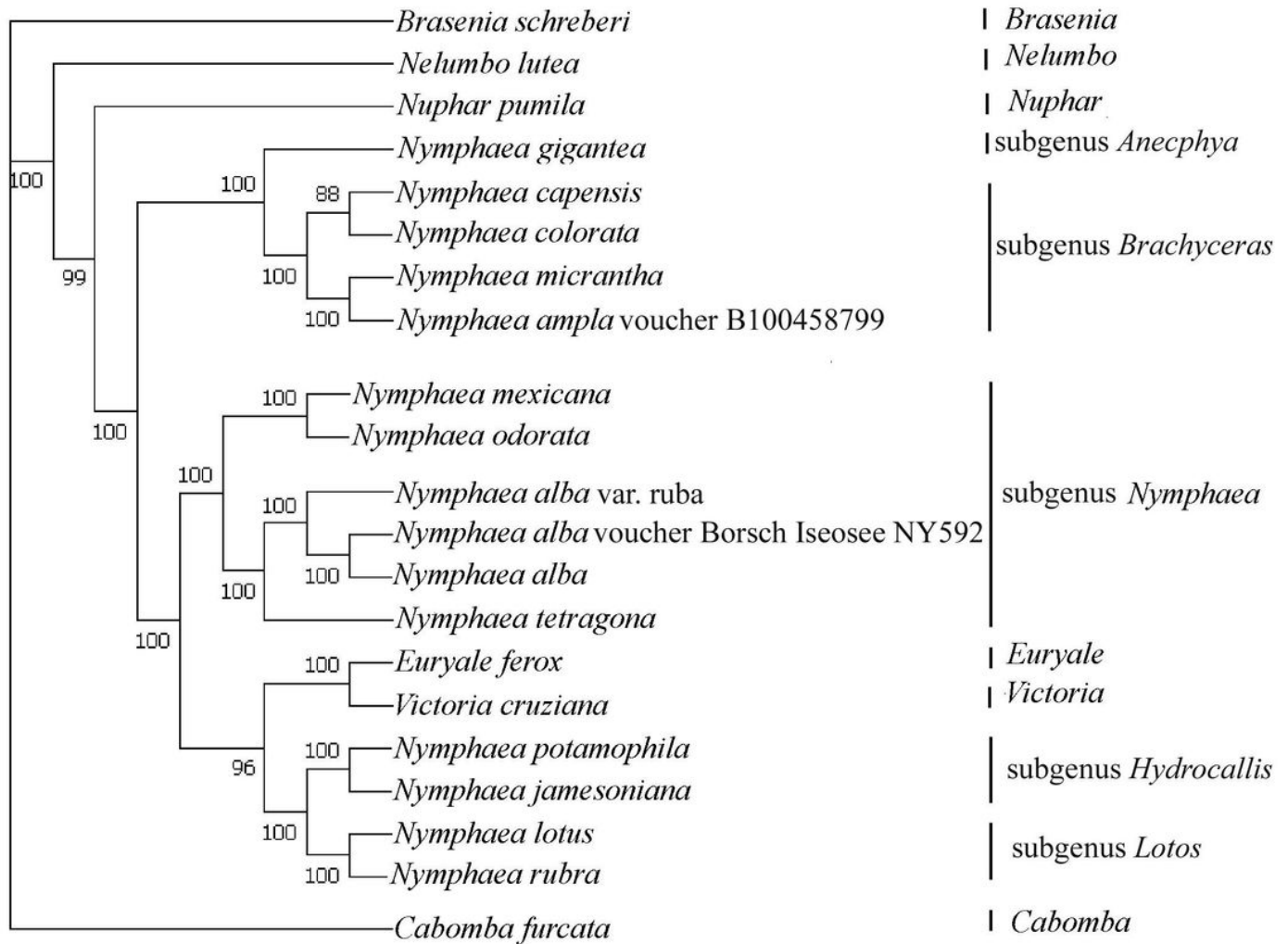
Figure 5

Five Nymphaea subgenera

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.xls

- Additionalfile2.xls

- Additionalfile5.xls

- Additionalfile4.xlt

- Additionalfile3.xls