SCIENCE CHINA
Life Sciences

• RESEARCH PAPER •

February 2013 Vol.56 No.2: 189–198
doi: 10.1007/s11427-012-4430-8

Complete chloroplast genome sequence of *Magnolia grandiflora* and comparative analysis with related species

LI XiWen[1,3†], GAO HuanHuan[1,2†], WANG YiTao[3], SONG JingYuan[1], HENRY Robert[4], WU HeZhen[2], HU ZhiGang[2], YAO Hui[1], LUO HongMei[1], LUO Kun[1], PAN HongLin[2] & CHEN ShiLin[1,5*]

[1]*Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences, Beijing 100193, China;*
[2]*Faculty of Pharmacy, Hubei University of Chinese Medicine, Wuhan 430065, China;*
[3]*Institute of Chinese Medical Science, University of Macao, Macao 999078, China;*
[4]*Queensland Alliance for Agriculture and Food Innovation, University of Queensland, Brisbane QLD 4072, Australia;*
[5]*Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing 100700, China*

Received July 13, 2012; accepted November 23, 2012; published online January 17, 2013

Magnolia grandiflora is an important medicinal, ornamental and horticultural plant species. The chloroplast (cp) genome of *M. grandiflora* was sequenced using a 454 sequencing platform and the genome structure was compared with other related species. The complete cp genome of *M. grandiflora* was 159623 bp in length and contained a pair of inverted repeats (IR) of 26563 bp separated by large and small single copy (LSC, SSC) regions of 87757 and 18740 bp, respectively. A total of 129 genes were successfully annotated, 18 of which included introns. The identity, number and GC content of *M. grandiflora* cp genes were similar to those of other Magnoliaceae species genomes. Analysis revealed 218 simple sequence repeat (SSR) loci, most composed of A or T, contributing to a bias in base composition. The types and abundances of repeat units in Magnoliaceae species were relatively conserved and these loci will be useful for developing *M. grandiflora* cp genome vectors. In addition, results indicated that the cp genome size in Magnoliaceae species and the position of the IR border were closely related to the length of the *ycf1* gene. Phylogenetic analyses based on 66 shared genes from 30 species using maximum parsimony (MP) and maximum likelihood (ML) methods provided strong support for the phylogenetic position of *Magnolia*. The availability of the complete cp genome sequence of *M. grandiflora* provides valuable information for breeding of desirable varieties, cp genetic engineering, developing useful molecular markers and phylogenetic analyses in Magnoliaceae.

intron, inverted repeats, SSR, phylogenetics

Citation: Li X W, Gao H H, Wang Y T, et al. Complete chloroplast genome sequence of *Magnolia grandiflora* and comparative analysis with related species. Sci China Life Sci, 2013, 56: 189–198, doi: 10.1007/s11427-012-4430-8

The chloroplast (cp) is an important semiautonomous organelle for plant photosynthesis, and is inherited in a maternal manner. Compared with the nuclear genome, the cp genome is less than one ten-thousandth of the size. The cp has multiple copies in each cell, which can allow high expression of genes targeted to the cp. Targeted integration or transgenes in the cp genome can avoid position effects and gene silencing [1]. The chloroplast has a moderate rate of nucleotide evolution, but shows a big difference in the rate of divergence between coding and non-coding regions. This makes the cp genome suitable for phylogenetic studies at different taxonomic levels [2]. Since the complete cp genome sequences of *Nicotiana tabacum* [3] and *Marchantia*

†Contributed equally to this work
*Corresponding author (email: slchen@implad.ac.cn)

polymorpha [4] were first sequenced in 1986, cp genomes have been widely used in cp genetic engineering, for developing useful molecular markers and for phylogenetic analyses [5–7]. Kanevski et al. [8] transferred the *rbcL* gene of sunflower into the cp genome of *N. tabacum* and greatly improved the photosynthesis rate of seedlings. Craig et al. [9] transformed the *Delta*9 dehydrogenase of wild potato into the cp genome of *N. tabacum* using the PEG method, and obtained highly frost-resistant plants. These results showed that cp genetic engineering had huge potential application in resistance breeding. Jansen et al. [10] constructed phylogenetic trees based on 81 shared genes of 64 cp genomes and recovered phylogenetic relationships between the main angiosperm groups, which explained the major relationships among the early evolutionary branches and solved taxonomic controversy over the basal groups among angiosperms.

Magnolia grandiflora Linn., with the popular Chinese names Guang Yulan and Yang Yulan, is a tree species originating from the southeast of North America. *M. grandiflora* has great economic value, as well as strong resistance to wind and toxic gases such as sulfur dioxide. It has beautiful flowers with a fragrant smell and is regarded as an important ornamental and horticultural species. Extracts from its leaves are used to reduce blood pressure and as a raw material in Chinese herbal medicines, including use as a substitute for *M. officinalis*. However, seedlings of *M. grandiflora* grow slowly and are prone to diseases and insect damage. In addition, young plants are not hardy at low temperatures (<−14°C). Therefore, *M. grandiflora* has a low survival rate when transplanted and subsequently supplies of its products fall short of demand. *M. grandiflora* belongs to a group of plants that are early basal species of the angiosperms. Taxonomic boundaries are not well defined between this genus and other genera of Magnoliaceae, especially in relation to characters such as internal structures and external shapes. Agreement on the phylogeny of the Magnoliaceae has not been possible using traditional methods such as morphology, anatomy, microscopy and genetic analysis [11–14]. Apart from *Liriodendron*, the other genera of Magnoliaceae have been incorporated into *Magnolia*. The relationships of *Magnolia* have been a focus of evolutionary and taxonomic studies [15]. Recent studies of *M. grandiflora* have focused on tissue culture, chemical composition and pharmacological effects [16–18], but research has failed to solve many issues such as the long growth period, low resistance to diseases, classification and conservation of biodiversity.

These problems would be expected to be solved by cp genetic engineering and phylogenetic analysis. The absence of a cp genome sequence for *M. grandiflora* posed an insurmountable barrier to current studies in these areas. We now report the cp genome structure of *M. grandiflora.* This supports the study of evolution in the early angiosperms. The complete genome sequence provides valuable genetic information for studies on photosynthetic mechanisms, cultivating new varieties with strong resistance to cold and insect damage, exploring phylogenetic relationships between species of Magnoliaceae, and can also offer basic knowledge for the cp genetic engineering of *M. grandiflora*.

1 Materials and methods

1.1 Plant materials

Fresh young leaves of *M. grandiflora* were harvested from Nanjing (E118°27′36″, N32°1′48″). The samples were identified by Lin YuLin, an Associate Professor of the Institute of Medicinal Plant Development (IMPLAD), and preserved at IMPLAD.

1.2 Extraction and sequencing

Total cpDNA was extracted from approximately 100 g leaves using a sucrose gradient centrifugation method that was improved by Li XiWen et al. [19]. The concentration of the DNA for each cp was estimated by measuring A_{260} with an ND-2000 spectrometer (Nanodrop technologies, Wilmington, DE, USA), and visual approximation was performed using gel electrophoresis. Pure cpDNA was sequenced using a 454/Roche FLX high-throughput sequencing platform.

1.3 Assembly and annotation

The Sff-file obtained was pre-processed, including the trimming of low-quality sequences. *De novo* assembly was performed using version 2.5 of the GS FLX system software. The position and direction of the contigs were identified using the cp genome sequence of *Liriodendron tulipifera* (NC_008326) as the reference sequence. The boundaries of IR-LSC and IR-SSC were confirmed using PCR amplification. The forward primer (F) sequence for LSC-IRb was CCTTCTCTCTTTCTCTCGCC and the reverse primer (R) was ATGAACCCTGTAGACCATCC. Other boundaries and their primer sequences were IRb-SSC (F-GCAGAATACCGTCGCCTAT, R-TACATTGCTCAAGTTGTGCC), SSC-IRa (F-CTGTGCCAAGGTTTCAGAC, R-AAACAGGAACAAGAGGCATC), and IRa-LSC (F-CAATGGAGCCGTAGACAGT, R-CATCAATCGTGCTAACCTTG).

The new complete cp genome sequence was annotated with the aid of DOGMA (http://dogma.ccbb.utexas.edu/) online. The position of each gene was determined using a blast method with the complete cp genome sequence of *L. tulipifera* as a reference sequence. Minor revisions were performed according to the start and stop codons.

1.4 Drawing a physical map of the cpDNA

A physical map of the chloroplast genome was produced by

exporting the sequence in GenBank format using the Sequin software and submitting the cp genome sequence of *M. grandiflora* to the GenomeVx website (http://wolfe.gen.tcd.ie/GenomeVx/).

1.5 Analysis of simple sequence repeats (SSR)

SSR loci were identified using MISA (http://pgrc.ipk-gatersleben.de/misa/) with the parameters set to eight repeat units (⩾8) for mononucleotide SSRs, four repeat units (⩾4) for dinucleotide, and three repeat units (⩾3) for trinucleotide, tetranucleotide pentanucleotide and hexanucleotide SSRs. We focused on perfect repeat sequences. The sequences of cyclic queues and reverse complements were regarded as the same SSR. For example, the repeat unit AAG is equal to AAG, AGA, GAA, CTT, TCT and TTC.

1.6 Phylogenetic tree

There were altogether 66 genes (Table S1) shared by all 30 cp genomes (Table S2). Each gene from each genomic sequence was modified manually by checking the start and stop codons and was aligned using ClustalW. The resulting gene alignments were assembled into a data matrix for each genome. Maximum Likelihood (ML) and Maximum Parsimony (MP) analyses were conducted using PAUP (Phylogenetic Analysis Using Parsimony) v. 4.0b10 (Swofford, USA) taking the cp genome sequence of *Cycas taitungensis* (NC_009618) as the outgroup. The appropriate model of evolution was determined using the Modeltest3.7 software.

2 Results and discussion

2.1 The cp gene features of *M. grandiflora*

The complete cp genome sequence of *M. grandiflora* was 159623 bp in length and contained a pair of inverted repeats (IR) of 26563 bp separated by large and small single copy (LSC, SSC) regions of 87757 bp and 18740 bp, respectively (Figure 1). A total of 129 genes were successfully annotated, including 37 tRNA, 8 rRNA and 84 protein coding genes (Table 1). Seven tRNAs and all rRNAs were located in IR regions. Protein coding regions accounted for 49% of the whole genome sequence, while rRNA, tRNA, and intergenic regions and introns accounted for 5.66%, 1.74%, and 43.6%, respectively. The GC-content of the whole cp genome sequence was 39.3%, and in the IR regions was 43%. This was a little higher than in the LSC and SSC regions (38% and 34% respectively). The size of the cp genome, GC content and gene content in *M. grandiflora* were found to be similar to those in *M. officinalis*, *M. kwangsiensis* and *L. tulipifera* (Table 2), which was consistent with the ex-

Table 1 Gene list for the *M. grandiflora* cp genome[a)]

Category for genes	Group of genes	Name of gene
Self replication	Ribosomal RNA genes	*rrn16*[b] *rrn23*[b] *rrn4.5*[b] *rrn5*[b]
	Transfer RNA genes	*trnA-UGC*[a,b] *trnC-GCA* *trnD-GUC* *trnE-UUC* *trnF-GAA* *trnfM-CAU* *trnG-GCC* *trnG-UCC*[a] *trnH-GUG* *trnI-CAU*[b] *trnI-GAU*[a,b] *trnK-UUU*[a] *trnL-CAA*[b] *trnL-UAA*[a] *trnL-UAG* *trnM-CAU* *trnN-GUU*[b] *trnP-UGG* *trnQ-UUG* *trnR-ACG*[b] *trnR-UCU* *trnS-GCU* *trnS-GGA* *trnS-UGA* *trnT-GGU* *trnT-UGU* *trnV-GAC*[b] *trnV-UAC*[a] *trnW-CCA* *trnY-GUA*
	Small subunit of ribosome	*rps2* *rps3* *rps4* *rps7*[b] *rps8* *rps11* *rps12*[a,b] *rps14* *rps15* *rps16*[a] *rps18* *rps19*
	Large subunit of ribosome	*rpl2*[a,b] *rpl14* *rpl16*[a] *rpl20* *rpl23*[b] *rpl32* *rpl33* *rpl36*
	DNA dependent RNA polymerase	*rpoA* *rpoB* *rpoC1*[a] *rpoC2*
	Translational initiation factor	*infA*
Genes for photosynthesis	Subunits of photosystem I	*psaA* *psaB* *psaC* *psaI* *psaJ*
	Subunits of photosystem II	*psbA* *psbB* *psbC* *psbD* *psbE* *psbF* *psbH* *psbI* *psbJ* *psbK* *psbL* *psbM* *psbN* *psbT* *psbZ*
	Subunits of cytochrome	*petA* *petB*[a] *petD*[a] *petG* *petL* *petN*
	Subunits of ATP synthase	*atpA* *atpB* *atpE* *atpF*[a] *atpH* *atpI*
	ATP-dependent protease subunit p gene	*clpP*[a]
	Large subunit of Rubisco	*rbcL*
	Subunits of NADH dehydrogenase	*ndhA*[a] *ndhB*[a,b] *ndhC* *ndhD* *ndhE* *ndhF* *ndhG* *ndhH* *ndhI* *ndhJ* *ndhK*
Other genes	Maturase	*matK*
	Envelop membrane protein	*cemA*
	Subunit of Acetyl-CoA-carboxylase	*accD*
	c-type cytochrome synthesis gene	*ccsA*
Genes of unknown function	Conserved open reading frames	*ycf1* *ycf2*[b] *ycf3*[a] *ycf4*

a) a, Genes containing introns; b, duplicated gene (genes present in the IR regions).

Table 2 Comparison of general features of the plastid genomes in Magnoliidae

Genome features	*M. grandiflora*	*M. officinalis*	*M. kwangsiensis*	*L. tulipifera*
Total Length (bp)	159623	160183	159667	159886
GC content (%)	39.30	39.22	39.26	39.16
LSC Length (bp)	87757	88210	88030	88150
SSC Length (bp)	18740	18843	18669	18964
IR Length (bp)	26563	26565	26484	26386
Total genes	129	126	129	129
Genes duplicated in IR	17	17	17	17
Protein genes	84	81	84	84
rRNA genes	8	8	8	8
tRNA genes	37	37	37	37

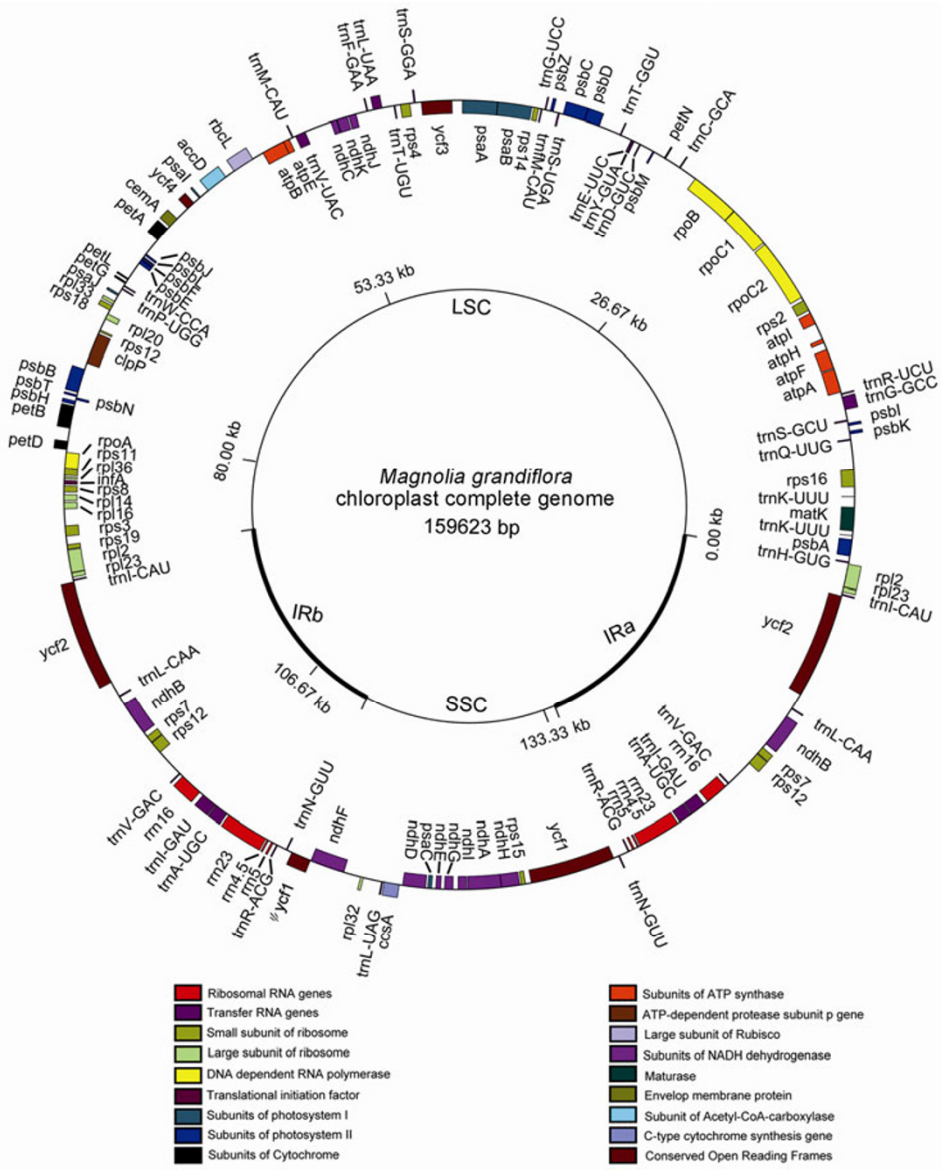

Figure 1 Representative map of the cp genome of *M. grandiflora*. The thick lines in the inner ring indicate the extent of the inverted repeats (IRa and IRb), which separate the genome into the large single-copy (LSC) and single-copy (SSC) regions. Genes on the outside of the map are transcribed in the clockwise direction and genes on the inside of the map are transcribed in the counterclockwise direction.

pectation that Magnoliaceae species evolve slowly [20].

The intron plays an important role in the regulation of gene expression. Some recent studies have found that many introns improve exogenous gene expression at specific positions and times, resulting in the expected agronomic characters. Therefore, introns can be a useful tool to improve transformation efficiency [21]. There were 18 intron-containing genes in the cp genome of *M. grandiflora*, of which three genes (*rps12*, *ycf3* and *clpP*) had two introns and the rest had one intron (Table 3). These introns are useful loci and may help to develop new varieties with strong resistance to many environmental stresses. Five genes were duplicated in the IRs, including *trnA-UGC*, *trnI-GAU*, *rps12*, *rpl2* and *ndhB*. *rps12* is a specific trans-splicing gene of which the 5′ end of the exon was located in LSC region and the 3′ end in the IR regions. This phenomenon was also found in other plastomes such as *Marchantia polymorpha* [22] and Korean ginseng [23]. It was reported that *rpl16* and *petD* had no introns in *M. kwangsiensis* and *L. tulipifera* but we found highly similar sequences in the corresponding positions by performing a blast search of the intron sequences of *rpl16* and *petD* in the cp genomes of *Calycanthus floridus* var. *glaucus*, *Drimys granadensis*, *Piper cenocladum* and *M. grandiflora*. We deduced that the absence of introns could be an error of annotation. In addition, Kuang et al. [24] reported that the shorter exon in *rpl16* was absent in *M. kwangsiensis* and *L. tulipifera.* The conflict between our results and previous studies needs to be solved with rigorous experimental validation. The absence of exons and introns has also played an important role in gene structure and differentiation of function [25] and provides more information to study the evolution of closely related species. The findings in this paper indicate that the sequence variations of *rpl16* might have a significant influence on phylogenetic studies in Magnoliidae and the study of functional changes related to the *rpl16* genes in different species.

2.2 The analysis of SSRs

Chloroplast simple sequence repeats (SSR) are effective molecular markers. They not only have the advantages of abundance, co-dominant inheritance and high repeatability, but also the characteristics of simple genomic structure, relatively conservative sequences and maternal inheritance, which makes them widely used in species identification and genetic analysis at individual and group levels [26,27].

A total of 218 SSR loci were present in the cp genome of *M. grandiflora*, including 91 mononucleotide, 44 dinucleotide, 72 trinucleotide, 9 tetranucleotide and 2 hexanucleotide repeat units. No pentanucleotide repeats were found, and there was an inverse relationship between the abundance and the lengths of repeat units. Among all of the cpSSRs, the repeat unit A/T was the most abundant repeat, followed by AG/CT, AAG/CTT, AT/AT, AAT/ATT and AAC/GTT. These repeat units accounted for 81.2% of the total SSRs (Table 4). In addition, mononucleotide, dinucleotide, and trinucleotide repeats were composed of A or T at a higher level; this contributed to a bias in base composition, which was consistent with the overall A-T richness (60.7%) of the cp genome. The bias may have a close relationship with the easier changes to A-T rather than G-C in the genome. A survey of SSRs in other species of Magnoliaceae was performed with the same parameters used in *M. grandiflora*, allowing a comparison of the distribution of repeat units. The results showed that the type and abundance of repeat units in Magnoliaceae species were quite conserved.

Table 3 Characteristics of genes including introns and exons in the cp genome of *M. grandiflora*[a)]

Gene	Exon I	Intron I	Exon II	Intron II	Exon III
trnA-UGC*	38	798	35		
trnG-UCC	24	767	48		
trnI-GAU*	42	936	35		
trnK-UUU	37	2491	35		
trnL-UAA	35	490	50		
trnV-UAC	37	584	39		
rps12*	114	–	231	537	30
rps16	42	829	246		
rpl2*	384	661	432		
rpl16	9	960	411		
atpF	144	706	411		
petB	6	784	642		
petD	8	653	525		
ndhA	540	1078	552		
ndhB*	755	703	775		
ycf3	153	731	228	734	126
clpP	246	631	291	778	69
rpoC1	432	734	1614		

a) Genes with an asterisk are located in the IR regions.

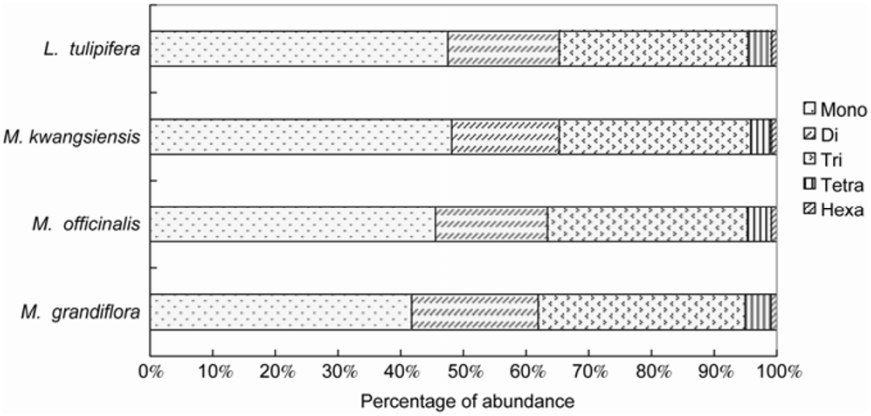

Figure 2 SSRs in the cp genomes of *M. grandiflora* and other species of Magnoliaceae. Mono represents mononucleotide, Di represents dinucleotide, Tri represents trinucleotide, Tetra represents tetranucleotide, and Hexa represents hexanucleotide repeats.

Table 4 Type and abundance of different SSR repeat units in *M. grandiflora*

SSR repeat units	SSR abundances	Percent abundance (%)
Mononucleotide		
A/T	88	96.7
C/G	3	3.3
Dinucleotide		
AC/GT	2	4.55
AG/CT	24	54.55
AT/AT	18	40.91
Trinucleotide		
AAC/GTT	11	15.28
AAG/CTT	20	27.78
AAT/ATT	16	22.22
ACC/GGT	4	5.56
ACT/AGT	2	2.78
AGC/CTG	8	11.11
AGG/CCT	5	6.94
ATC/ATG	6	8.33
Tetranucleotide		
AAAT/ATTT	3	33.33
AACT/AGTT	1	11.11
AATC/ATTG	1	11.11
AATG/ATTC	2	22.22
AATT/AATT	1	11.11
ACAT/ATGT	1	11.11
Hexanucleotide		
AATACT/AGTATT	2	100

These loci should be useful for discovery of universal SSR markers for the Magnoliaceae.

2.3 Comparison of IR boundaries in Magnoliidae

Differences in cp genome size are mainly caused by the contraction and expansion of the IR regions [28]. Comparison of the IR boundary among six species from five orders of Magnoliaceae (Figure 3) showed that the size of the IR regions has a positive relationship with the length of the complete cp genome sequence (except in *L. tulipifera*). The cp genome size of *P. cenocladum* was the largest among the six species, followed by *D. granadensis*, *M. grandiflora*, *L. tulipifera*, *Chloranthus spicatus* and *C. floridus* var. *glaucus* (Table S3). Correlation analysis (Table S4) indicated, except for *C. floridus* var. *glaucus* (Figure S1), that the length of the IR regions had a negative correlation with that of the pseudogene *ycf1* (*ψycf1*) (R^2=0.81, P<0.05). *C. floridus* var. *glaucus* had the shortest IR region and its *ψycf1* was also shortest (only 266 bp). We also found a non-normal distribution of *ψycf1* length between *C. floridus* var. *glaucus* and the other species. The *ψycf1* length value of *C. floridus* var. *glaucus* belonged to an extremely unusual data category in Spearman analysis. In addition, apart from the contraction of *ψycf1*, the stretching of intergenic regions between *rps19* and *rpl2* (up to 1553 bp) altered the length of the IRs in *C. floridus* var. *glaucus*. The length of the IR and its corresponding *ψycf1* are listed in Table S3. The *ψycf1* of *P. cenocladum* was only 927 bp in length but had the longest IR compared with other species in the Magnoliidae. The length of *ψycf1-ndhF* reached up to 1106 bp, which created an obvious expansion of the IR regions. Overlaps were detected between *ψycf1* and *ndhF* in *D. granadensis* and *C. spicatus* with lengths of 72 and 25 bp, respectively. The stretching of the IR in *D. granadensis* produced a duplicated *trnH* gene. The *trnH* gene was partly located in the IR region of *C. spicatus*. We also found a close relationship between the changes in IR regions and the length between the *rps19* and IR boundary. The presence of *ψtrnH* and a duplicated *rpl2* was closely related to this phenomenon. The changes in IR-SSC were mainly reflected in the length of *ycf1*. *P. cenocladum* had the shortest *ycf1* gene sequence completely located within the IR, which was the largest among the six species. The IR of *C. floridus* var. *glaucus* was the shortest and had the smallest percentage of the length of *ycf1*. However, we determined that the changes in *ycf1* length and position had no significant correlation with that of the IR (P>0.05). The position of the *trnH* gene located at the junction of IRa/LSC shifted obviously but irregularly.

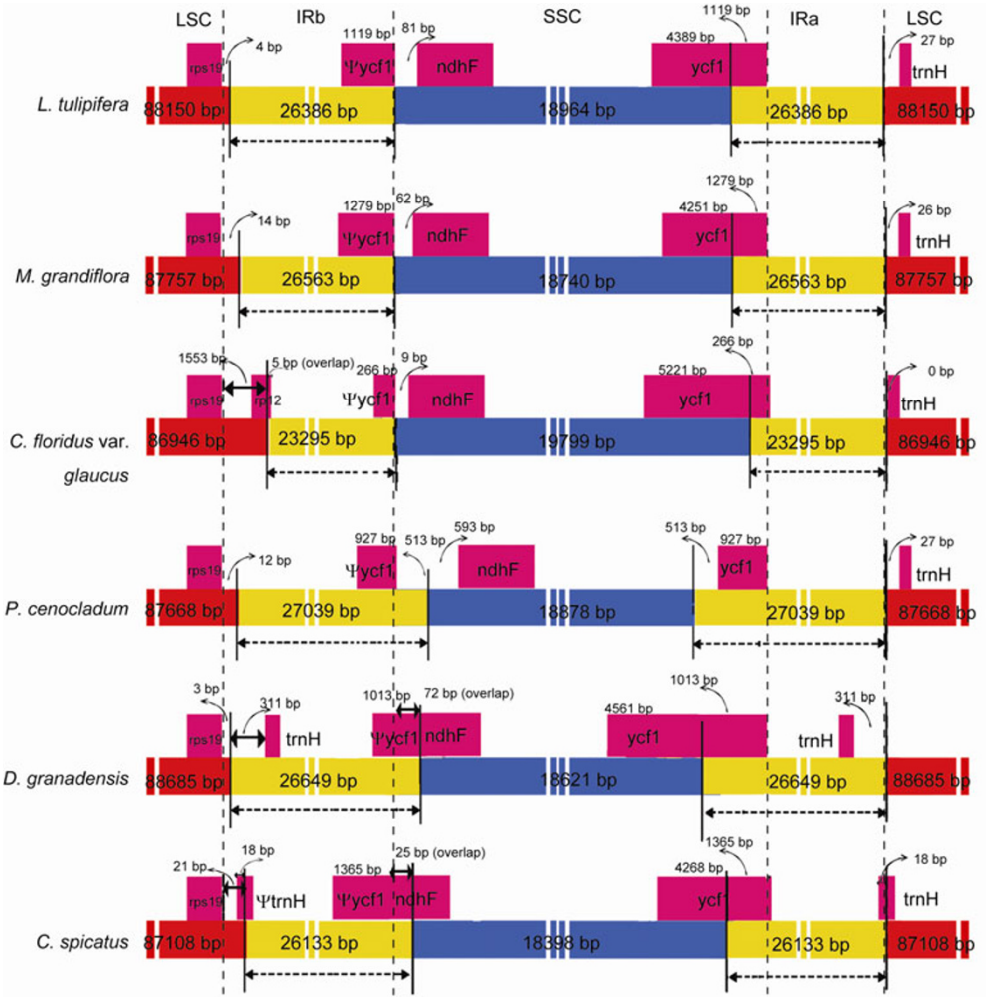

Figure 3 Comparison of the LSC, IR and SSC border regions among six cp genomes. This figure shows six cp genome sequences from Magnoliidae, and each sequence contains a pair of inverted repeats (IRb and IRa) separated by large and small single copy (LSC, SSC) regions. The IR regions have the same color. Genes near the borders of the IRs are labeled, including *rps19*, *ndhF*, *trnH* and *ycf1*. The distances between these genes and the borders are marked with curved arrows. The overlap indicates that the *ndhF* and *ψycf1* genes share the same sequences.

2.4 Phylogenetic analysis of *Magnolia* species using cp genome sequences

Phylogenetic analysis was performed using maximum likelihood (ML) (Figure 4) and maximum parsimony (MP) (Figure 5) methods on a 30-taxon 66-gene data matrix with 52869 aligned nucleotide positions, but when the gaps were excluded there were 47786 characters. Model test analysis showed that the best model was GTR+I+G. Bootstrap analysis indicated that 26 of the 27 nodes were supported by values ⩾90% and 19 of these had a bootstrap value of 100% in the ML tree. MP analysis also generated a single tree with a high proportion of highly supported nodes, with 88.5% (23/26) of the ingroup nodes resolved with bootstrap values ⩾90%, 16 of which were 100%. Although the taxon sampling was inadequate and we could not make a deeper phylogenetic analysis of Magnoliidae species, this was the first analysis of the phylogenetic position of *Magnolia* using cp genome data. The results of MP and ML analyses showed that *Magnolia* was sister to *Liriodendron*, and the two genera contributed to a later diverging lineage of the basal angiosperms. ML analysis determined that Magnoliales+Lauralesis was sister to Piperales+Canellales, which was consistent with the reports of Jansen et al. [10] and Cai et al. [20]. In addition, the phylogenetic tree with cp genome sequences recovered Monocotyledoneae as sister to Eudicotyledoneae with a high support value and this group was sister to Chloranthaceae+Magnoliidae; this result was identical to previous studies published by Michael et al. [29] and Jansen et al. [10]. Compared with the ML tree, the MP tree supported Laurales+Piperales as sister to Canellales and followed by Magnoliidae. However, the support value between Laurales and Piperales was lower, only 52%. Although there have been many phylogenetic studies between Magnoliaceae species using genes or intergenetic regions, the support values have been very low. Single gene and intergenetic region sequences have a low frequency of parallel evolution in genera of the Magnoliaceae [30–33],

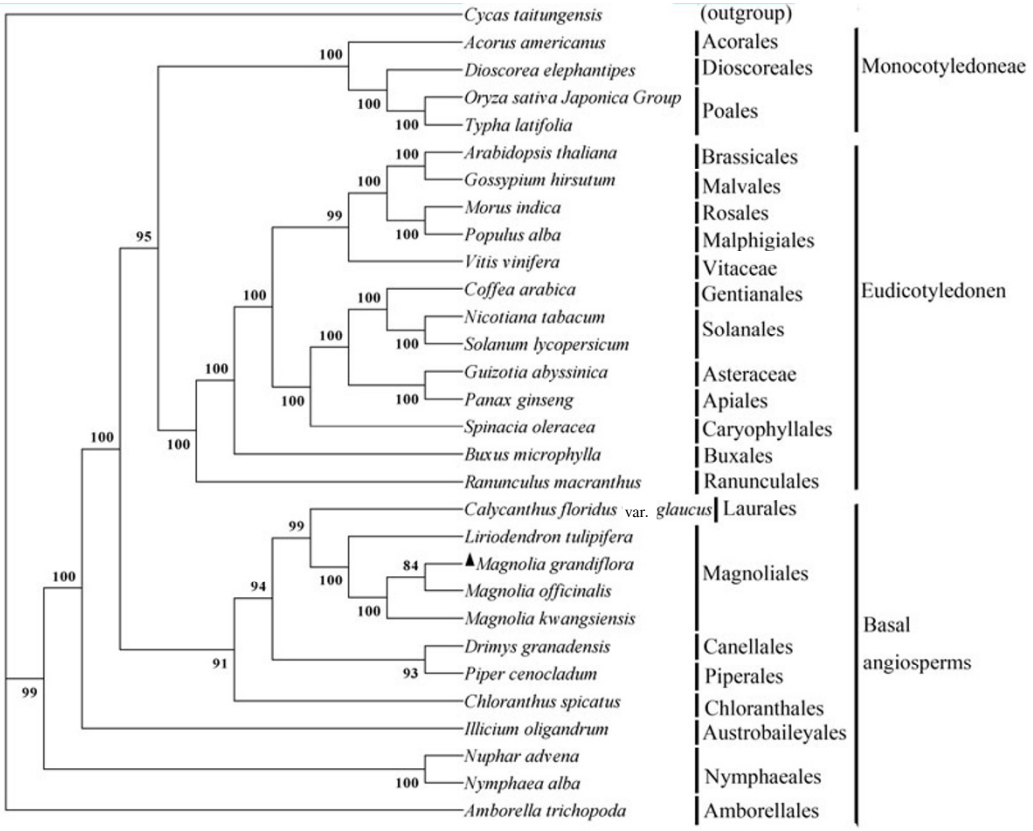

Figure 4 ML tree (−ln*L*=353897.8125) of 30 species based on a 66-gene data set.

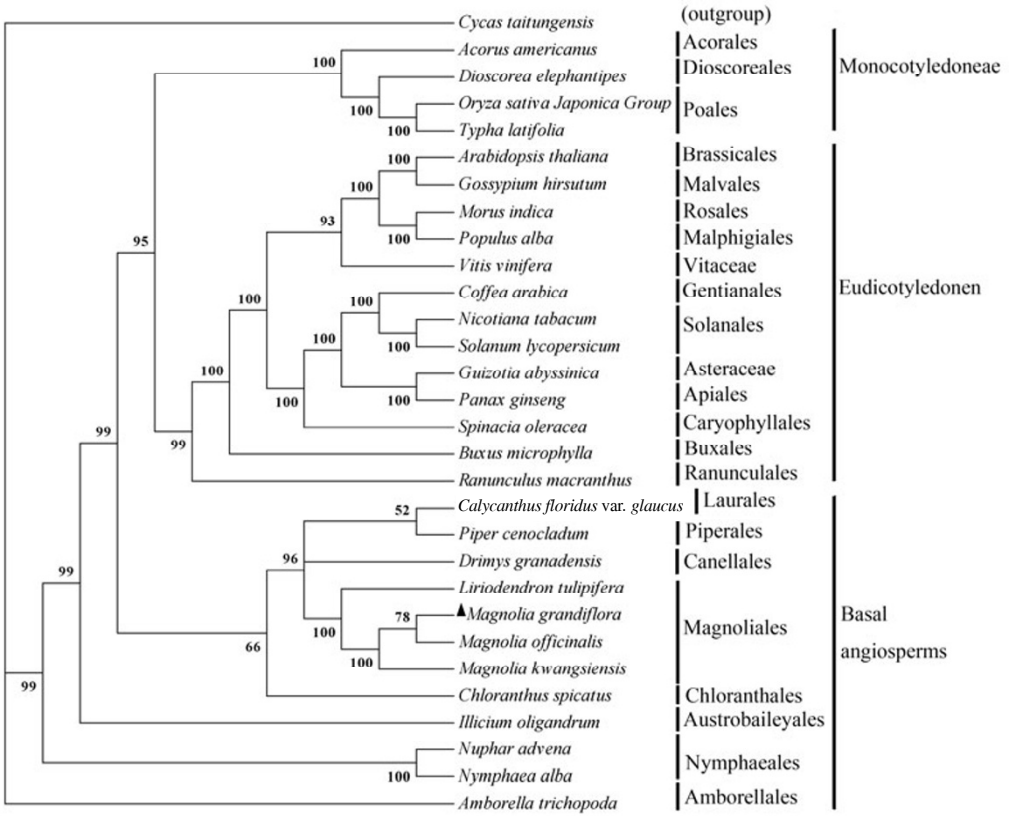

Figure 5 MP tree of 30 species based on a 66-gene data set.

therefore they are only suitable for phylogenetic studies at higher taxonomic levels (above genus) [34,35]. Because the cp genome evolves slowly in Magnoliaceae species, deeper phylogenetic research has to depend on cp genome data at lower taxonomic levels.

This study investigated the evolutionary position of *Magnolia* among the basal angiosperms and the phylogenetic relationship between four orders of Magnoliaceae based on previous phylogenetic studies on basal angiosperms using cp genome sequences. Our study provides new data for deeper analysis of phylogeny within the basal angiosperms. Existing cp genome sequences of Magnoliaceae remain sporadic and we need more cp genome data to determine the position of different evolutionary branches. We believe that more cp genomes of Magnoliid species will be completely sequenced with the sequencing cost decreasing and the development of new assembly technologies. Their complicated phylogenetic relationships will be finally defined according to systematic analysis with cp genome data.

3 Conclusion

With the further advancement of high-throughput sequencing technologies and bioinformatics, we can make deeper studies of the cp genome and learn more about its origin and structure, which will help us to explore new molecular markers and study the phylogenetic relationships of plant species. Moreover, studying the cp genome can help advance cp genetic engineering including molecular breeding and genetic transformation. *M. grandiflora* is a valuable ornamental and medicinal plant. Sequencing its complete cp genome and analyzing its structure can offer basal genetic information on photosynthetic regulation, enable development of varieties with strong resistances and assist plant identification. The chloroplast genome sequence can also help us find new ways to solve diverse problems for this species such as slow growth, insect pests and phylogenetic issues.

This work was supported by the National Natural Science Foundation of China (30970307, 81130069) and the Innovation Research Team of the University of Ministry of Education of China (IRT1150).

1 Verma D, Daniell H. Chloroplast vector systems for biotechnology applications. Plant Physiol, 2007, 145: 1129–1143
2 Clegg M T, Gaut B S, Learn G H, et al. Rates and patterns of chloroplast DNA evolution. Proc Natl Acad Sci USA, 1994, 91: 6795–6801
3 Shinozaki K, Ohme M, Tanaka M, et al. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J, 1986, 5: 2043–2049
4 Ohyama K, Fukuzawa H, Kohchi T, et al. Chloroplast gene organization deduced from complete sequence of Liverwort *Marchantia polymorpha* chloroplast DNA. Nature, 1986, 322: 572–574
5 Terada R, Urawa H, Inagaki Y, et al. Efficient gene targeting by homologous recombination in rice. Nat Biotechnol, 2002, 20: 1030–1034
6 Kane N, Sveinsson S, Dempewolf H, et al. Ultra-barcoding in cacao (Theobroma spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. Am J Bot, 2012, 99: 320–329
7 Zhang Y J, Ma P F, Li D Z. High-Throughput Sequencing of Six Bamboo Chloroplast Genomes: Phylogenetic Implications for Temperate Woody Bamboos (Poaceae: Bambusoideae). PLoS One, 2011, 6: e20596
8 Kanevski I, Maliga P, Rhoades D F, et al. Plastome engineering of ribulose-1, 5-bisphosphate carboxylase/oxygenase in tobacco to form a sunflower large subunit and tobacco small subunit hybrid. Plant Physiol, 1999, 119: 133–142
9 Craig W, Lenzi P, Scotti N, et al. Transplastomic tobacco plants expressing a fatty acid desaturase gene exhibit altered fatty acid profiles and improved cold tolerance. Transgenic Res, 2008, 17: 769–782
10 Jansen K J, Zhengqiu C, Linda A R, et al. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci USA, 2007, 104: 19369–19374
11 Xu F, Rudall P J. Comparative floral anatomy and ontogeny in Magnoliaceae. Plant Syst Evol, 2006, 258: 1–15
12 Yamada T, Imaichi R, et al. The outer integument and funicular outgrowth complex in the ovule of *Magnolia grandiflora* (Magnoliaceae). J Plant Res, 2003, 116: 189–198
13 Kim S, Park C W, Kim Y D, et al. Phylogenetic relationships in family Magnoliaceae inferred from *ndhF* sequences. Am J Bot, 2001, 88: 717–728
14 Sauquet H, Doyle J A, Scharaschkin T, et al. Phylogenetic analysis of Magnoliales and Myristicaceae based on multiple data sets: implications for character evolution. Bot J Linn Soc, 2003, 142: 125–186
15 Fu D L. Notes on Yulania Spach. J Wuhan Bot Res, 2001, 19: 191–198
16 Wang Q, Wang Z Z, Li Y L. Study on tissue culture of *Magnolia grandiflora* L.. Northwest Pharm J, 2001, 16: 11–13
17 Lee S, Chappell J. Biochemical and genomic characterization of terpene synthases in *Magnolia grandiflora.* Plant Physiol, 2008, 147: 1017–1033
18 Wang Z G, Wu C Q, Wang C Y, et al. Effect of *Magnolia grandifore* oil on lipid metabolism in hyperlipoidemic rat. Chin Trad Patent Med, 2010, 32: 1679–1682
19 Li X W, Hu Z G, Lin X H, et al. High-throughput pyrosequencing of the complete chloroplast genome of *Magnolia officinalis* and its application in species identification. Acta Pharm Sin, 2012, 47: 124–130
20 Cai Z Q, Penaflor C, Kuehl J V, et al. Complete plastid genome sequence of *Drimys*, *Liriodendron*, and *Piper*: implications for the phylogenetic relationships of magnoliids. BMC Evol Biol, 2006, 6: 77
21 Xu J W, Feng D J, Song G S, et al. The first introns in rice EPSP synthase enhance exogenous gene expression. Science China Ser C-Life Sci, 2003, 33: 224–230
22 Fukuzawa H, Kohchi T, Shirai H, et al. Coding sequences for chloroplast ribosomal protein S12 from the liverwort *Marchantia polymorpha*, are separated far apart on the different DNA Strand. FEBS Lett, 1986, 198: 11–15
23 Kim K J, Lee H L. Complete chloroplast genome sequences from Korean Ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. DNA Res, 2004, 11: 247–261
24 Kuang D Y, Wu H, Wang Y L, et al. Complete chloroplast genome sequence of *Magnolia kwangsiensis* (Magnoliaceae): implication for DNA barcoding and population genetics. Genome, 2011, 54: 663–673
25 Xu G X, Guo C C, Shan H Y, et al. Divergence of duplicate genes in exon–intron structure. Proc Natl Acad Sci USA, 2012, 109: 1187–1192
26 Kaundun S S, Matsumoto S. Heterologous nuclear and chloroplast microsatellite amplification and variation in tea *Camellia sinensis*.

Genome, 2002, 45: 1041–1048
27 Jiao Y, Jia H M, Li X W, et al. Development of simple sequence repeat (SSR) markers from a genome survey of Chinese Bayberry (*Myrica rubra*). BMC Genomics, 2012, 13: 201
28 Ravi V, Khurana J P, Tyagi A K, et al. An update on chloroplast genomes. Plant Syst Evol, 2008, 271: 101–122
29 Michael J M, Soltis, P S, Bell C D, et al. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. Proc Natl Acad Sci USA, 2010, 107: 4623–4628
30 Azuma H, Thien L B, Kawano S. Molecular phylogeny of *Magnolia* (Magnoliaceae) inferred from cpDNA sequences and evolutionary divergence of the floral scents. J Plant Res, 1999, 112: 291–306
31 Azuma H, Thien L B, Kawano S. Molecular phylogeny of *Magnolia* based on chloroplast DNA sequence data (*trnK* intron, *psbA-trnH* and *atpB-rbcL* intergenic spacer regions) and floral scent chemistry. In: Liu Y H, Fan H M, eds. Proceedings of the International Symposium on the Family Magnoliaceae. Beijing: Science Press, 2000. 219–227
32 Azuma H, Garcia-Franco J G, Rico-Gray V, et al. Molecular phylogeny of the Magnoliaceae: the biogeography of tropical and temperate disjunctions. Am J Bot, 2001, 88: 2275–2285
33 Ueda K, Yamashita J, Tamura M N. Molecular phylogeny of the Magnoliaceae. In: Liu Y H, Fan H M, eds. Proceedings of the International Symposium on the Family Magnoliaceae. Beijing: Science Press, 2000. 205–209
34 Wang Y L, Zhang S Z, Cui T C. The utility of *trnL* intron and *trnL-trnF* IGS in phylogenetic analysis of Magnoliaceae. Acta Bot Boreali-Occident Sin, 2003, 23: 247–252
35 Wang Y L, Li Y, Zhang S Z, et al. The utility of *matk* gene in the phylogenetic analysis of the genus *Magnolia*. Acta Phytotaxon Sin, 2006: 135–147

Supporting Information

Table S1 Sequences from Genbank used for phylogenetic analysis

Table S2 The total shared 66 genes used in the phylogenetic analysis

Table S3 Comparison of the cp genome lengths, IR regions and pseudogenes among Magnoliidae species

Table S4 Spearman correlation analysis between the IR region and *ψycf1* length among Magnoliidae species

Figure S1 The linear-distribution analysis of length between the IR regions and *ψycf1*.

The supporting information is available online at life.scichina.com and www.springerlink.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.