热带亚热带植物学报 2015, 23(6): 601 ~ 513
Journal of Tropical and Subtropical Botany

药用植物华重楼(黑药花科)叶绿体全基因组研究

李晓娟[1,2], 杨振艳[1], 黄玉玲[1,2], 纪运恒[1*]
(1. 中国科学院昆明植物研究所东亚植物多样性与生物地理学重点实验室,昆明 650201;2. 中国科学院大学,北京 100049)

摘要: 为探究华重楼(*Paris polyphylla* var. *chinensis*)的叶绿体基因组特征,利用叶绿体系统发育基因组学方法,对华重楼与其它百合目植物的叶绿体全基因组进行了比较。结果表明,华重楼的叶绿体全基因组长158307 bp,由 4 个区组成,包括 2 个反向重复区(IRA和IRB, 27473 bp)、1 个小单拷贝区(SSC, 18175 bp)和 1 个大单拷贝区(LSC, 85187 bp)。其叶绿体基因组有115个基因,包括81个编码蛋白质基因、30 个转运RNA基因和 4 个核糖体RNA基因。11 种百合目植物的叶绿体全基因组的基因组成和基因顺序相似。华重楼的 *cemA*基因是假基因,其起始密码子后有多聚核苷酸 poly(A)及CA双核苷酸重复序列,编码序列中出现多个终止密码子 , 且与北重楼(*Paris verticillata*)的*cemA*编码序列中的终止密码子位置不同。因此,华重楼叶绿体基因组比较保守;*cemA*结构及假基因化现象可能具有重要的进化与系统发育信息,其编码序列中的终止密码子可以区分华重楼和北重楼。

关键词: 叶绿体全基因组;华重楼;黑药花科;百合目;*cemA* 假基因化
doi: 10.11926/j.issn.1005–3395.2015.06.001

Complete Chloroplast Genome of the Medicinal Plant *Paris polyphylla* var. *chinensis* (Melanthiaceae)

LI Xiao-juan[1,2], YANG Zhen-yan[1], HUANG Yu-ling[1,2], JI Yun-heng[1*]
(1. *Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming institute of Botany, Chinese Academy of Sciences,* Kunming 650201, China; 2. *University of Chinese Academy of Sciences,* Beijing 100049, China)

Abstract: In order to understand the characters of chloroplast genome (cp genome) in *Paris polyphylla* var. *chinensis*, the chloroplast genome (cp genome) of *P. polyphylla* var. *chinensis* was compared with those of 10 species within Liliales by using phylogenomics methods based on complete chloroplast genomes. The results showed that the cp genome of *P. polyphylla* var. *chinensis* was 158307 bp in length and display a typical quadripartite structure including two inverted repeat regions (IRA and IRB, 27473 bp), one small single-copy region (SSC, 18175 bp) and one large single-copy region (LSC, 85187 bp). It contained 115 unique genes, including 81 protein-coding genes, 30 tRNAs and 4 rRNAs. The genome structure, gene contents and arrangement of 10 Liliales species cp genomes were very similar. The *cemA* gene of *P. polyphylla* var. *chinensis* was pseudogene with poly(A) and CA SSR patterns after the start codon, and the loci of premature stop codons are different from those of *Paris veticillata*. In conclusion, the cp genome of *P. polyphylla* var. *chinensis* was conservative. The *cemA* structure and pseudogenization might play an important role in the evolution and phylogeny, and the location of the stop codons in *cemA* was useful for distinguishing *P. polyphylla* var. *chinensis* from *P. veticillata*.

Key words: Complete chloroplast genome; *Paris polyphylla* var. *chinensis*; Melanthiaceae; Liliales; *cemA* pseudogenization

Received: 2015–05–12 **Accepted:** 2015–05–27
This work was supported by the National Natural Science Foundation of China (Grant No. 30670132).
* Corresponding author. E-mail: jiyh@mail.kib.ac.cn

The chloroplast (cp) genome in plants is about 120–160 kp in size. Because of its small genome size and high copy numbers per cell, sequencing the complete cpDNA genome is much more amenable than the nuclear genome of plants[1–3]. Besides with highly conserved gene content and order across plant species, cp genome DNA sequences were reported that it can provide useful information to elucidate phylogenetic relationships among plant taxa[4–5]. In addition, cp genome has a relatively lower intraspecific and higher interspecific divergence than nuclear genome, so that species identification can be confirmed easily depending on whether a gene exist in either of two species[6–7]. At this time, the number of chloroplast genomes of green plants uploaded to NCBI (http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=2759&opt=plastid#pageTop) has risen to 644, of which almost 496 angiosperms (including 139 monocots) have been available until March 4th, 2015.

Paris polyphylla var. *chinensis* (Franch.) Hara. is a perennial herb in the tribe Parideae of the family Melanthiaceae, which is widespread in subtropical China[8–10]. The plant (named "Chonglou" in Chinese) is a traditional Chinese medicinal herb whose dried rhizome is the main source of raw material for some famous prepared Chinese medicines such as "Yunnan Baiyao" and "Gong xue ning Capsules", which are used as haemostatic, antalgic and antipyretic[11–12]. More than 50 chemical compounds have been isolated and identified from *P. polyphylla* var. *chinensis* to date[13–15]. Modern pharmacological research has demonstrated that steroid saponins (Diosgenin and Pennogenin glycosides) are responsible for these biological activities[15–17].

Although *P. polyphylla* var. *chinensis* has a great economic and medicinal importance, no genomic work on the plant has yet been performed. Recently, three plastid genomes have been sequenced and confirmed in the Melanthiaceae family, including *Veratrum patulum* O. Loes[18] (tribe Melanthieae), *Chionographis japonica* (Willd) Maxim[19] (tribe Chionographideae) and its congener, *Paris verticillata* M. Bieb[20]. The *trnI_CAU* triplication and *cemA* pseudogenization were found in the complete cp genomic sequence of *P. verticillata*[20], which proposed a hypothesis that these patterns would be useful for understanding the phylogeny and evolution of these species by comparing the plastid genome features among Liliales. Here we reported the complete cp genome of *P. polyphylla* var. *chinensis*, and contrasted it with those of *P. veticillata* and other Liliales taxa. These data will be helpful to understand the phylogenetic relationships and evolutionary mechanism within Parideae clade (Melanthiaceae), and to provide useful molecular information for further research on this important medicinal plant.

1 Material and methods

1.1 Taxon sampling, cpDNA extraction and sequencing

In general, there are mainly three ways for obtaining cpDNA genome sequence: the first, cpDNA was isolated using low pH medium with high salt method[21]; the second, amplifying cpDNA using short range PCR primers for Sanger DNA sequencing[22–23]; the third, isolating total genomic DNA, constructing DNA library and sequencing utilizing next-generation sequencing[24–25]. However, these ways need large sums of fresh leaves samples for sequencing and cpDNA sequence obtained may be with considerable gaps. To avoid these problems, we used the method proposed by Yang et al[26] to amplify the whole cp genome of *P. polyphylla* var. *chinensis* with nine universal primers. The healthy, actively growing fresh leaves were collected from *P. polyphylla* var. *chinensis* cultivated in the green house of Kunming Institute of Botany, Chinese Academy of Sciences. Total genome was extracted from about 100 mg of these clean, fresh leaves using CTAB method. The complete chloroplast genome of *P. polyphylla* var. *chinensis* was amplified by Takara PrimeSTAR GXL DNA polymerase and nine universal pairs of primers[26]. Purified PCR products were mixed and then broken into 200–500 bp fragments and set paired-end libraries according to the

manufacturer's manual (Illumina). The libraries were sequenced (2×100 bp) by Illumina Hiseq 2000.

1.2 Genome assembly

Before assembling short raw reads of *P. polyphylla* var. *chinensis* into contigs, sequence data were filtered using NGS QC Tool Kit[27] for high quality (cut-off value for percentage of read length=80, cut-off value for PHRED quality score=30). Then the filtered reads were conducted to contigs with the *de novo* sequence assembly software, CLC Genomics Workbench[28] V. 7 on Windows7 64bits server, and the word size was set to 64, while the minimum contig lengh was set to 1 kb. One cp genome which was highly similar to that of *P. polyphylla* var. *chinensis* was obtained after contigs were aligned using the Basic Local Alignment Search Tool (http://blast.ncbi.nlm.nih.gov/) with default parameters. We arranged the aligned contigs using the highly similar genome sequence identified in the BLAST search as references, and joined the contigs according the orders of contigs. At this time, contigs were assembled into a genome sequence with some gaps. In order to get one complete chloroplast genome, we also assembled reads into contigs using SOAPdenovo2[29], on linux server, set the kmers to 81. We joined the contigs into another incomplete genome sequence following the same steps. Last, we aligned the two incomplete genome sequences and filled gaps. The complete chloroplast genome sequence of *P. polyphylla* var. *chinensis* was assembled.

1.3 Genome annotation, drawing and comparison

The *P. polyphylla* var. *chinensis* cpDNA genome was annotated using the program DOGMA (http://phylocluster.biosci.utexas.edu/dogma/) and start and stop codons were adjusted using Geneious 7.0 software[30]. The tRNA genes were identified and corrected by the tRNAscan-SE (http://selab.janelia.org/tRNAscan-SE/). The pseudogenes were defined in terms of terminal codons found in the middle of protein genes coding sequence[31]. The gene introns were showed in gene annotation tables which Geneious[30] displayed. Then we used the annotated chloroplast genome genome file to draw gene map utilizing OrganellarGenomeDRAW (http://ogdraw.mpimpgolm.mpg.de/index.shtml). We compared the chloroplast genomic characteristic of *P. polyphylla* var. *chinense* with 47 species chosen from some orders of monocot plants included Liliales, Zingiberales, Poales, Arecales, Asparagales, Dioscoreales, Petrosaviales, Alismatales and Acorales, which data were downloaded from NCBI (Table 1). The 48 taxa genome sequences of *cemA* were aligned by MUSCLE in Geneious plugins, and the alignment result was corrected manually.

1.4 Sequence divergence and phylogenetic analysis

The 48 complete plastid sequences representing the nine orders of monocots (Table 1) were downloaded from NCBI Organelle Genome Resources database. We extracted 79 protein-coding genes (all protein genes except *ycf15* and *ycf68*) from 48 species sequences. The question marks were substituted for missing genes of some taxa, and missing genes sequence were aligned by MUSCEL, respectively. These alignment results were modified manually. Pairwise sequence divergences were calculated with Kimura's two parameter model using the software MEGA 6[32].

Because the *accD, ycf15, ycf68, ycf1, ycf2* are missing in some monocot lineages, we constructed the aligned matrix of 76 protein genes without those five genes using Phyutility.jar[33], and used this matrix to reconstruct phylogenetic tree with GTR+I+G substitution model computed by Modeltest, which is included in the PUAP GUI 2011 program. The Markov chain Monte Carlo (MCMC) algorithm was run for 1000000 generations with trees sampled every 100 generations for each data partition. The first 25% of trees from all runs were discarded as burn-in, and the remaining trees were used to construct majority-rule consensus tree. We chose *Acorus americanus* (Raf.) Raf. and *A. calamus* L. as outgroups. The phylogenetic tree was conducted by FigTree V1.4 program.

Table 1 Accession number of complete chloroplast genomes

Order	Family	Species	Accession number
Acorales	Acoraceae	*Acorus americanus*	NC_010093
		A. calamus	NC_007407
Alismatales	Hydrocharitaceae	*Elodea canadensis*	NC_018541
		Najas flexilis	NC_021936
	Araceae	*Colocasia esculenta*	NC_016753
		Spirodela polyrhiza	NC_015891
		Lemna minor	DQ400350
		Wolffia australiana	NC_015899
		Wolffiella lingulata	NC_015894
Petrosaviales	Petrosaviaceae	*Petrosavia stellaris*	NC_023356
Dioscoreales	Dioscoreaceae	*Dioscorea elephantipes*	EF380353
		D. rotundata	KJ490011
Liliales	Alstroemeriaceae	*Luzuriaga radicans*	NC_025333
		Bomarea edulis	KM233641
		Alstroemeria aurea	KC968976
	Smilacaeae	*Smilax china*	HM536959
	Liliacceae	*Lilium longiflorum*	KC968977
		Fritillaria cirrhosa	NC_024728
		F. hupehensis	NC_024736
	Melanthiaceae	*Veratrum patulum*	NC_022715
		Chionographis japonica	KF951065
		Paris verticillata	KJ433485
Zingiberales	Musaceae	*Musa textilis*	NC_022926
	Heliconiaceae	*Heliconia collinsiana*	NC_020362
	Strelitziaceae	*Ravenala madagascariensis*	NC_022927
	Marantaceae	*Maranta leuconeura*	KF601571
	Zingiberaceae	*Zingiber spectabile*	NC_020363
		Curcuma roscoeana	NC_022928
Poales	Bromeliaceae	*Ananas comosus*	NC_026220
	Typhaceae	*Typha latifolia*	NC_013823
		Anomochloa marantoidea	NC_014062
		Pharus lappulaceus	NC_023245
	Poaceae	*Leersia tisserantii*	NC_016677
		Yushania lebigata	NC_024725
		Lolium perenne	AM777385
Arecales	Arecaceae	*Bismarckia nobilis*	NC_020366
		Phoenix dactylifera	NC_013991
		Cocos nucifera	KF285453
		Elaeis guineensis	NC_017602
		Calamus caryotoides	JX088663
Asparagales	Orchidaceae	*Neottia nidus-avis*	NC_016471
		Dendrobium officinale	KC771275
		Phalaenopsis equestris	NC_017609
		Oncidium	NC_014056
		Cymbidium aloifolium	NC_021429
	Amaryllidaceae	*Allium cepa*	KF728079
	Asparagaceae	*Eustrephus latifolius*	KM233639

2 Results

2.1 Genome features

The complete cp genome of *P. polyphylla* var. *chinensis* is 158307 base pairs (bp) in length, which exhibits a quadripartite structure, consisting of a pair of IRs (27473 bp) separated by the large single-copy region (LSC, 85187 bp) and small single-copy region (SSC, 18175 bp) (Fig. 1). The overall CG content is 62.8%. The GC content of the IRs, LSC and SSC regions are 41.7%, 35.5% and 31.4%, respectively.

The cp genome of *P. polyphylla* var. *chinensis* encodes 137 predicted functional genes, of which 115 are unique, including 81 protein-coding genes, 4 rRNAs, and 30 tRNAs (Table 2). Ten protein-coding, 8 tRNA and 4rRNA genes are duplicated in the IR regions, however, only a part of *ycf1* gene is duplicated in the junction between IRB and SSC regions. The LSC region includes 60 protein-coding and 21 tRNA genes, whereas the SSC region contains 12 protein-coding (including partial ycf1 gene) and 1 tRNA genes. There are 18 intron-containing genes (12 protein-coding and 6 tRNAs) within the cp genome of *P. polyphylla* var. *chinensis*. Among them, 9 protein-coding genes and 6 tRNAs contain one intron, and 3 protein-coding genes (*clpP, ycf3, rps12*) have two introns. The *cemA*, *ycf15* and *ycf68* are pseudogenes judging by the presence of several terminal codons in these coding genes regions. In the *cemA* gene sequence, ploy(A) (8 bp) sequence and a small single repeat (SSR) CA unit including 6 CA copies are found within the coding regions (Fig. 2).

2.2 Comparison with other cp genomes in Liliales

The basic cp genomic features of *P. polyphylla* var. *chinensis* and those of nine species from other families within the Liliales order were compared, including *P. verticillata*, *C. japonica* and *V. patulum* (Melanthiaceae), *Fritillaria hupehensis* P. K. Hsiao & K. C. Hsia.[34] and *Lilium longiflorum* Thunb.[35] (Liliaceae), *Smilax china* L.[36] (Smilacaceae) and *Alstroemeria aurea* Graham.[34] and *Bomarea edulis* (Tussac) Herb.[37] (Alstroemeriaceae). The genome size of *P. polyphylla* var. *chinensis* is the largest of the Liliales cp genome. This variation in sequence length is mainly attributed to the difference in the length of the LSC region (Table 3). The AT and CG content ratio of *P. polyphylla* var. *chinensis* (37.2% and 62.8%, respectively) are similar to those of other species within the Liliales. The gene content and arrangement are similar within the Liliales lineages, except that *rps16* is deleted completely in *C. japonica*[24] but partially in *V. patulum*[18], *infA* lost in *A. aurea*[34] and *S. china*[36], and *ycf15* was absent from *A. aurea*[34], The gene content and arrangement are similar within the Liliales lineage. The IRB/SSC boundary is the incomplete duplication of *ycf1* in all species examined, whereas the IRA/LSC junction expands to *rps19* (*P. polyphylla* var. *chinensis, L. longiflorum*[35] and *A. aurea*[34]), full *trnH_GUG* (*V. patulum*[18], *F. hupehensis*[34] and *B. edulis*[37]), part of *rpl22* (*S. china*[36]) and part of *rps3* (*C. japonica*[19] and *P. verticillata*[20]). The length of intergenic spacer between *rpl23* and *ycf2* which contains *trnI-CAU* are varied among Liliales taxa. *P. polyphylla* var. *chinensis* cpDNA genome has the longest IGS length (636 bp) among all Liliales species (Table 3).

2.3 Sequence divergence of protein genes

We calculated the average pairwise sequence distance of 79 protein-coding genes among 48 monocots taxa. The results showed that 36 genes (45.57%) have an average sequence distance more than 0.10. The fourteen most divergent genes (*ycf1*, *rbl16, matK, rbl22, clpP, rbl32, rps16, ndhA, ndhF, ccsA rps11, ndhD, accD, infA*) exhibits that the average distance is higher than 0.15. The highest average sequence distance is observed in *ycf1* (0.23), which is located at the IR/SSC boundary and shows a fast evolutionary trend in monocots. The seven most conserve gene (*psbE, psbF, rps7, rps12, ndhB, rbl2* and *psbL*) possess the average sequence distance less than 0.05 (Table 4).

2.4 Phylogenetic analysis

To identify the phylogenetic position of *P. poly-*

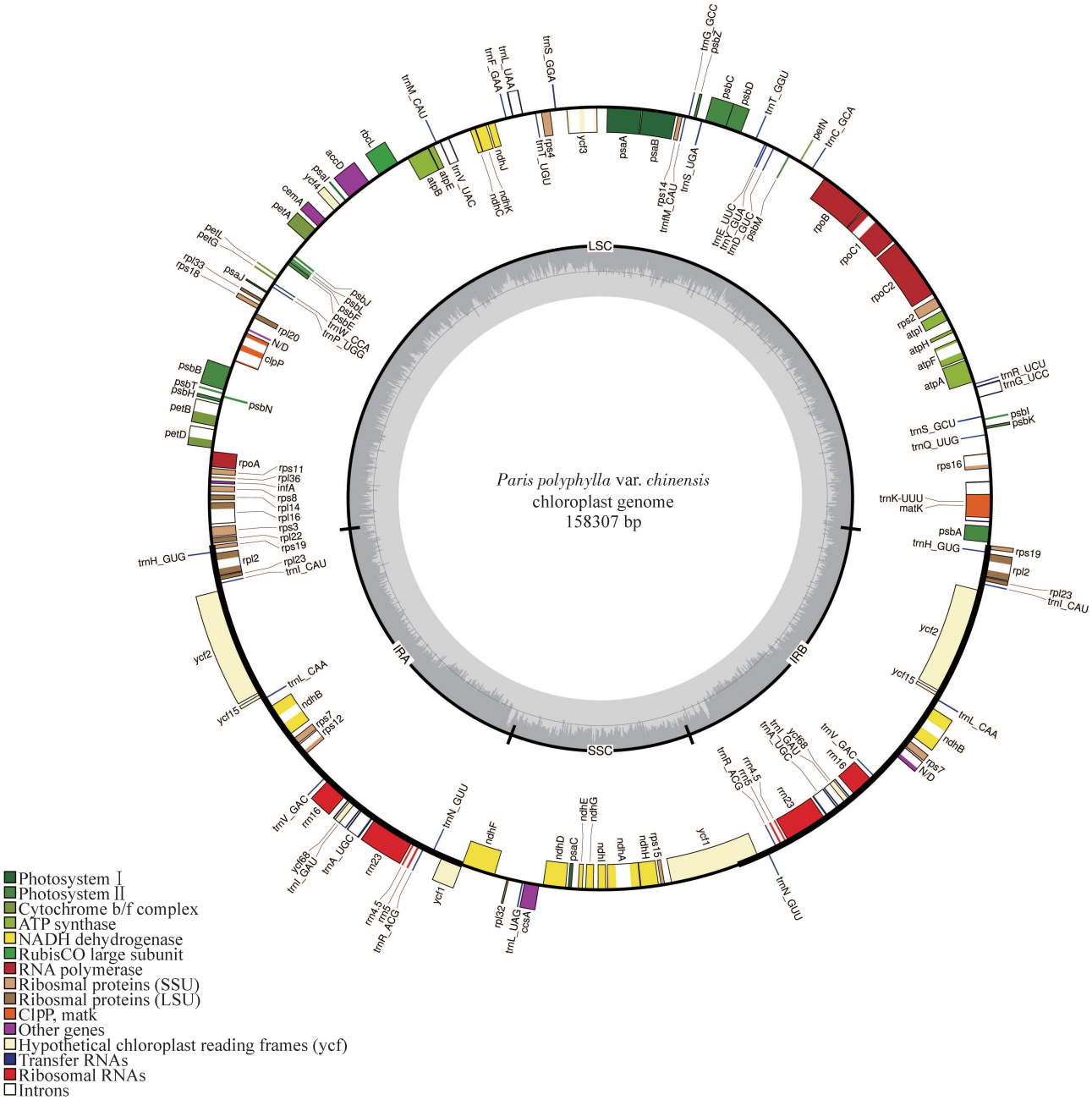

Fig. 1 Map of *Paris polyphylla* var. *chinensis* complete chloroplast genome. Transcribed counterclockwise are shown outside of outer circle, whereas transcribed clockwise are shown inside. Thick lines in outer circle indicate IR regions.

phylla var. *chinensis*, we carried out multiple sequence alignments by using 76 protein-coding genes in the cp genomes for 48 monocot taxa to generate a matrix of 85506 bp. Phylogenetic relationships in Melanthiaceae, Liliales, and other orders among monocots were reconstructed with MrBayes analysis (Fig. 3). The results indicated that Liliales is a monophyletic group [Bayesian posterior probabilities (BPP)=100%]. Within Liliales, Melanthiaceae is monophyletic (BPP=100%), which is sister to Smilacaceae (*S. china*) and Liliaceae (*L. longiflorum, F. cirrhosa, F. hupehensis*). Also, Alstroemeriaceae (*A. aurea*) is sister to Colchicaceae (*Cohchicum autrumnale, Gloriosa superba*). The genus *Paris* (BPP=100%) is monophyletic and sister to other taxa within Melanthiaceae.

Table 2 Gene contents of complete chloroplast genome of *Paris polyphylla* var. *chinensis*

Gene type	Genes
Photosystem I	*psaA, psaB, psaC, psaI, psaJ, ycf3***, *ycf4*
Photosystem II	*psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ psbK, psbL, psbM, psbN, psbT, psbZ*
Cytochrome	*petA, petB**, *petD**, *petG, petL, petN*
ATP synthase	*atpA, atpB, atpE, atpF**, *atpH, atpI*
Rubisco	*rbcL*
NADH dehydrogenase	*ndhA**, *ndhB** × 2, *ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*
Ribosomal protein (large subunit)	*rpl2** × 2, *rpl14, rpl16**, *rpl20, rpl22, rpl23* × 2, *rpl32, rpl33, rpl36*
Riosomal protein (small subunit)	*rps2, rps3, rps4, rps7* × 2, *rps8, rps11, rps12*** × 2, *rps14, rps15, rps16**, *rps18, rps19* × 2
ATP-dependent protease	*clpP***
Cytochrome c biogenesis	*ccsA*
Membrane protein	*cemA*
Maturase	*matK*
Other protein gene	*infA*
Proteins of unknown function	*ycf1* × 2, *ycf2* × 2, *ycf15* × 2, *ycf68* × 2
Ribosomal RNAs	*rrn23* × 2, *rrn16* × 2, *rrn5* × 2, *rrn4.5* × 2
Transfer RNAs	*trnA_UGC** × 2, *trnH_GUG* × 2, *trnR_ACG* × 2, *trnI_GAU** × 2, *trnI_CAU* × 2, *trnC_GCA, trnL_CAA* × 2, *trnV_GAC* × 2, *trnN_GUU* × 2, *trnD_GUC, trnE_UUC, trnF_GAA, trnG_UCC**, *trnG_UCC, trnK_UUU**, *trnL_UAA**, *trnV_UAC**, *trnL_UAG, trnM_CAU, trnfM_CAU, trnP_UGG, trnQ_UUG, trnR_UCU, trnS_GCU, trnS_UGA, trnS_GGA, trnT_GGU, trnT_UGU, trnW_CCA, trnY_GUA*
RNA polymerase	*rpoA, rpoB, rpoC1**, *rpoC2*

×2: Two gene copies in IR regions; *: With one intron; **: With two introns.

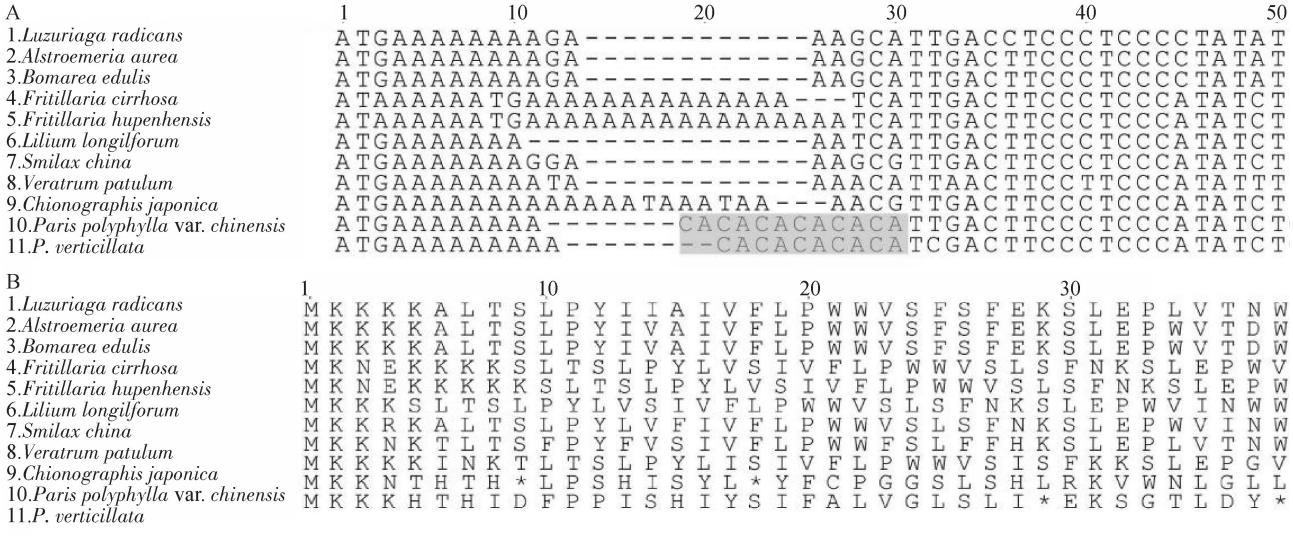

Fig. 2 Alignment of partial cemA sequences among 11 species of Liliales. A: Partial nucleotide sequences of *cemA*, gray box show SSR (CA); B: Partial amino acid sequences of cemA, asterisks indicate the stop condon.

3 Discussions

3.1 cpDNA genome features

The gene contents and arrangement were similar among Liliales as previously reported. However, there were some difference in genes loss, gene structure and pseudogenes in Liliales taxa, and some changes were typical among Liliales species. For instance, *rps16* was deleted completely in *C. japonica*[19], but partially in *V. patulum*[18]. The *infA* gene lost in *A. aurea*[34]

Table 3 Characteristics of chloroplast genomes among Liliales

	Paris polyphylla var. *chinensis*	*Paris verticillata*	*Chionographia japonica*	*Veratrum patulum*	*Fritillaria hupehensis*
Family	Melanthiaceae	Melanthiaceae	Melanthiaceae	Melanthiaceae	Liliaceae
Accession number		KJ433485	KF951065	KF437397	KF712486
Protein-coding genes	81	81	80	81	81
tRNAs	30	30	30	30	30
rRNAs	4	4	4	4	4
Length (bp)	158307	157379	154646	153699	152145
LSC (bp)	85187	82726	81653	83372	81898
SSC (bp)	18175	17907	18195	17607	17553
IRs (bp)	27473	28373	27399	26360	26347
AT content (%)	62.8	62.4	62.3	62.3	62.9
GC content (%)	37.2	37.6	37.7	37.7	37.1
IR/SSC junction	*ycf1*-like	*ycf1*-like	*ycf1*-like	*ycf1*-like	*ycf1*-like
IR/LSC junction	*rps19*	*rps3*-like	*rps3*-like	*trnH_GUG-rps19* IGS	*trnH_GUG-rps19* IGS
LSC GC content (%)	35.5	36	36	35.7	34.8
SSC GC content (%)	31.4	31.1	31.4	31.4	30.5
IR GC content (%)	41.7	42	42.5	42.9	42.5
Length of IGS (*rpl23-ycf2*)(bp)	636	591	303	305	307
	Lilium longiflorum	*Smilax china*	*Alstroemeria aurea*	*Bomarea edulis*	
Family	Liliaceae	Smilacaceae	Alstroemeriaceae	Alstroemeriaceae	
Accession number	KC968977	HM536959	KC968976	KM233641	
Protein-coding genes	81	81	79	80	
tRNAs	30	30	30	30	
rRNAs	4	4	4	4	
Length (bp)	152793	157878	155510	154925	
LSC (bp)	82230	84608	84241	84094	
SSC (bp)	17523	18536	17867	17699	
IRs (bp)	26520	27367	26701	26566	
AT content (%)	62.98	62.75	62.74	61.8	
GC content (%)	37.02	37.25	37.26	38.2	
IR/SSC junction	*ycf1*-like	*ycf1*-like	*ycf1*-like	*ycf1*-like	
IR/LSC junction	*rps19*-like	*rpl22*-like	*rps19*-like	*trnH_GUG-rps19* IGS	
LSC GC content (%)	34.8	35.2	36.2	36.4	
SSC GC content (%)	30.8	31.4	31.8	32.2	
IR GC content (%)	42.4	42.4	40.3	40.3	
Length of IGS (*rpl23-ycf2*)(bp)	308	308	308	307	

and *S. china*[36], while it existed as a pseudogene in the *V. patulum*[18], *B. edulis*[37] and *L. longiflorum*[35] because there were internal stop codons within the coding regions of this gene. The *accD* gene was pseudogene in *S. china*[36], whereas it had function within other species of Liliales. The gene loss and pseudogenization could be unique evolutionary events in those species.

Because of the presence of several stop codons, the *cemA* is pseudogene, and the gene is with poly(A) sequence and SSR of CA after the start codon in the cp genome of *P. polyphylla* var. *chinensis* as well as in *P. verticillata*[25]. However, the locations of stop codons in *cemA* are different between the cp

Table 4 Sequence divergence of protein genes

No.	Gene	Mean sequence distance	Standard error	Region	No.	Gene	Mean sequence distance	Standard error	Region
1	*ycf1*	0.23203	0.01059	IR	41	*rbL14*	0.09637	0.01633	LSC
2	*rbL16*	0.19853	0.01499	LSC	42	*rps18*	0.09584	0.01786	LSC
3	*matK*	0.19165	0.01289	LSC	43	*ndhk*	0.09547	0.01171	LSC
4	*rbL22*	0.18773	0.02473	LSC	44	*rbL36*	0.09479	0.03008	LSC
5	*clpP*	0.18683	0.01317	LSC	45	*rps14*	0.09387	0.01794	LSC
6	*rbL32*	0.17464	0.03499	SSC	46	*atpA*	0.09227	0.00777	LSC
7	*rps16*	0.17059	0.01485	LSC	47	*psbI*	0.09194	0.03019	LSC
8	*ndhA*	0.16579	0.01091	SSC	48	*rpoB*	0.09089	0.00538	LSC
9	*ndhF*	0.16052	0.00926	IR	49	*psaC*	0.08658	0.01925	SSC
10	*ccsA*	0.15776	0.01411	SSC	50	*rbcL*	0.08635	0.00804	LSC
11	*rps11*	0.15721	0.02079	LSC	51	*psaI*	0.08381	0.02821	SSC
12	*ndhD*	0.15337	0.01239	SSC	52	*atpB*	0.08203	0.00839	LSC
13	*accD*	0.15177	0.01125	LSC	53	*petA*	0.08031	0.00945	LSC
14	*infA*	0.15029	0.02688	LSC	54	*petL*	0.08027	0.02954	LSC
15	*rps15*	0.14918	0.02454	SSC	55	*psbJ*	0.07322	0.02484	LSC
16	*atpF*	0.14464	0.01152	LSC	56	*psbB*	0.07098	0.00707	SSC
17	*rbL33*	0.13968	0.03189	LSC	57	*petG*	0.07021	0.02542	LSC
18	*rbL20*	0.13798	0.02018	LSC	58	*psbC*	0.06873	0.00708	SSC
19	*rps3*	0.13509	0.01494	LSC	59	*atpI*	0.06811	0.00964	LSC
20	*rps8*	0.13195	0.01904	LSC	60	*psbM*	0.06476	0.02546	LSC
21	*rpoC2*	0.13129	0.00581	LSC	61	*psaA*	0.06379	0.00562	LSC
22	*ndhG*	0.12583	0.01715	SSC	62	*psaB*	0.06371	0.00582	LSC
23	*psbK*	0.12471	0.02701	LSC	63	*psbZ*	0.06255	0.01791	LSC
24	*cemA*	0.12293	0.01419	LSC	64	*psbT*	0.06172	0.02421	LSC
25	*rpoC1*	0.11994	0.00721	LSC	65	*ycf2*	0.06102	0.00389	IR
26	*rps19*	0.11922	0.02117	IR	66	*psbD*	0.05992	0.00761	LSC
27	*ndhC*	0.11635	0.01944	LSC	67	*petN*	0.05889	0.02628	LSC
28	*ycf3*	0.11574	0.00823	LSC	68	*atpH*	0.05837	0.01541	LSC
29	*rpoA*	0.11552	0.01122	LSC	69	*psbA*	0.05813	0.00727	SSC
30	*psaJ*	0.11477	0.03141	SSC	70	*psbN*	0.05773	0.02098	LSC
31	*atpE*	0.11308	0.01693	LSC	71	*psbE*	0.04825	0.01389	LSC
32	*ndhI*	0.11087	0.01662	SSC	72	*psbF*	0.04687	0.01954	LSC
33	*ndhH*	0.10877	0.01139	SSC	73	*rps7*	0.04038	0.00857	IR
34	*ndhE*	0.10764	0.01999	SSC	74	*rbL23*	0.03338	0.00974	IR
35	*petD*	0.10743	0.01088	LSC	75	*rps2*	0.03274	0.01299	LSC
36	*psbH*	0.10627	0.02248	LSC	76	*rps12*	0.03245	0.00634	LSC and IR (mainly in IR)
37	*ycf4*	0.09992	0.01363	LSC	77	*ndhB*	0.03089	0.00366	IR
38	*petB*	0.09797	0.01103	LSC	78	*rbl12*	0.03061	0.00451	IR
39	*ndhJ*	0.09708	0.01525	LSC	79	*psbL*	0.03004	0.01515	LSC
40	*rps4*	0.09639	0.01278	LSC					

genomes of *P. polyphylla* var. *chinensis* and *P. verticillata* (Fig. 2), which can be used to distinguish *P. polyphylla* var. *chinensis* and *P. verticillata*. The *cemA* encoding product was found in the inner envelope membrane of chloroplasts[38], which could be essential to CO_2 uptake in *Synechocystis*[39]. The *cemA* gene was found disappeared in the cp genome of two saprophytic monocots, e.g. *Neottia nidus-avi* (L.) Rich.[40], and *Petrosavia stellaris* Becc.[41], which could be interpreted by the dependence on the host plant. However, all species in the genus *Paris* are autotrophic, further research is needed to clarify the impact of *cemA* pseudogenization in the genus.

Gene duplication in the cp genome occurs mainly

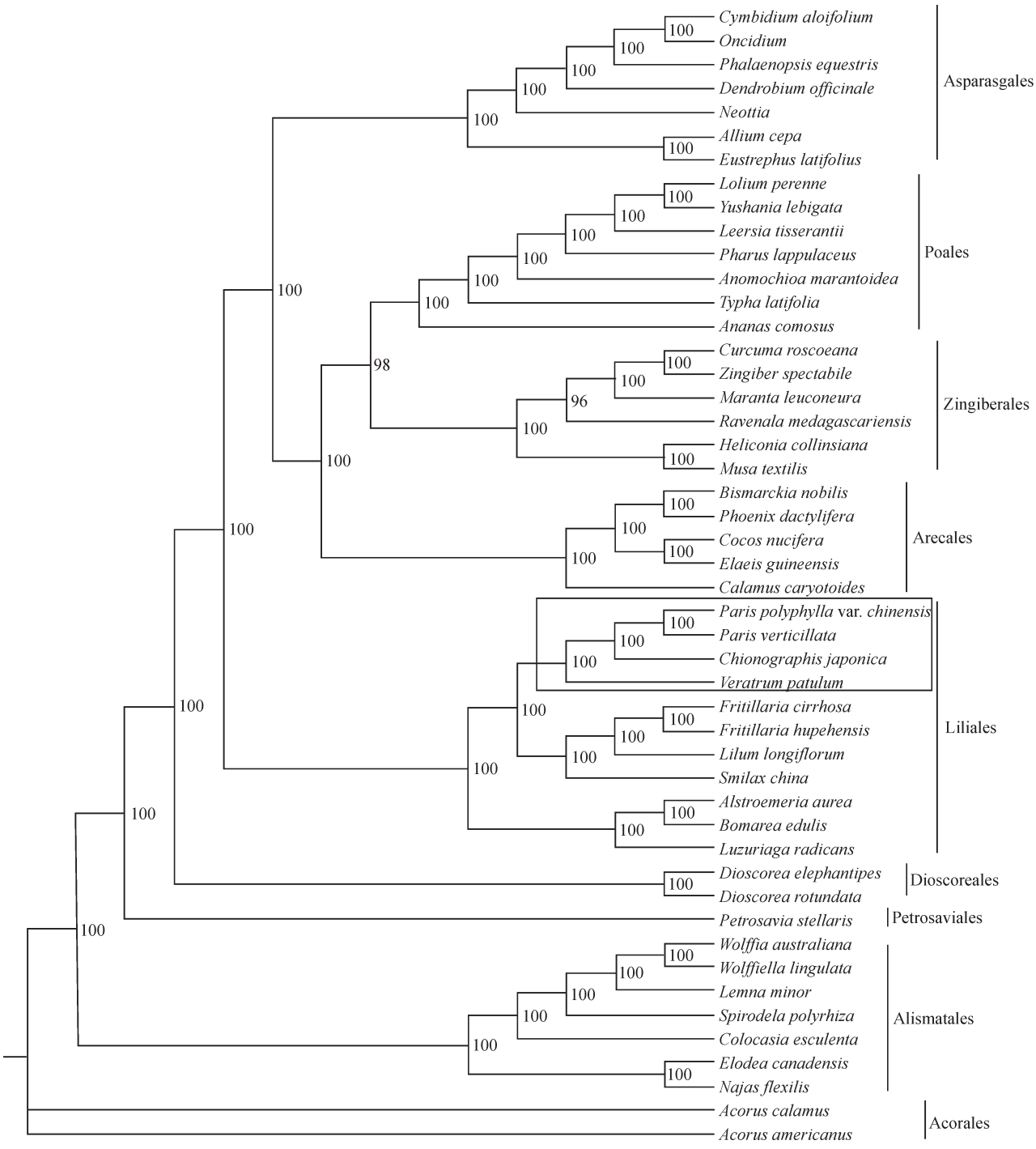

Fig. 3 Mrbayes tree inferred from 76 protein coding genes from 48 taxa. Bayesian posterior probabilities (BPP) are shown on the right of branches. The box displays the family Melanthiaceae taxa.

within the IR regions because of the IR region expansion[42] and most duplication genes were tRNAs[43]. The triplication of *trnI_CAU* has been reported in the cp genome of *P. verticillata*[20] but was not found in previously examined cp genomes of Melanthiaceae and other families in the Liliales, which was proposed to be unique to the tribe Parideae of Melanthiaceae. However, our data revealed that the triplication of *trnI_CAU* does not occur in *P. polyphylla* var. *chinensis*. *P. verticillata* and *P. polyphylla* var. *chinensis* belong two different subgenera of *Paris*[44]. Therefore, it is likely that the triplication of *trnI_CAU* occurs only in the subgenus *Paris* rather than the subgenus *Daiswa*. This difference may provide information to explore the infrageneric relationships within the genus *Paris*.

3.2 Phylogenetic relationships

Chloroplast genomes provide rich sources of phylogenetic information to elucidate the evolutionary relationships among angiosperms[4,45]. The order relationships among monocots and family relationships within Liliales defined in this study were identical to those delineations in previous studies[10,46]. However, the generic relationships among the family Melanthiaceae have not been satisfactory resolved inferred from our data due to the limited taxa sampling. *Paris* is a temperate genus with 27 species distributed in the Eurasia[8,44]. Previous phylogenetic studies employing one or several genes or DNA regions placed the genus in the family Melanthiaceae or the family Trilliaceae[9–10,47]. Our phylogeny based on 76 chloroplast protein genes placed *Paris* in the family Melanthiaceae with 100% BBP, which well supports the treatments of APG III[10], Fuse et al[9], and Zomlefer et al[47].

Acknowledgments We are very grateful to Dr. Hong-tao Li, Zheng-shan He, Li-e Yang, Cui-xian Peng, Jian-jun Jin of Kunming Institute of Botany for their help during processing our data. We thank Molecular Biology Experimental Center, Kunming Institute of Botany for experiments and materials.

References

[1] Li W M, Ruf S, Bock R. Constancy of organellar genome copy numbers during leaf development and senescence in higher plants [J]. Mol Genet Genom, 2006, 275(2): 185–192. doi: 10.1007/s00438-005-0075-7

[2] Mcneal J R, Leebens-Mack J H, Arumuganathan K, et al. Using partial genomic fosmid libraries for sequencing complete organellar genomes [J]. Biotechniques, 2006, 41(1): 69–73.

[3] Wakasugi T, Tsudzuki T, Sugiura M. The genomics of land plant chloroplasts: Gene content and alteration of genomic information by RNA editing [J]. Photosynth Res, 2001, 70(1): 107–118. doi: 10.1023/A:1013892009589

[4] Jansen R K, Cai Z Q, Raubeson L A, et al. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns [J]. Proc Natl Acad Sci USA, 2007, 104(49): 19369–19374. doi: 10.1073/pnas.0709121104

[5] Odintsova M S, Yurina N P. Chloroplast genomics of land plants and algae [M]// Biotechnological Applications of Photosynthetic Proteins: Biochips, Biosensors and Biodevices. US: Springer, 2006: 57–72. doi: 10.1007/978-0-387-36672-2_6

[6] Luo H W, Sun Z Y, Arndt W, et al. Gene order phylogeny and the evolution of methanogens [J]. PLoS One, 2009, 4(6): e6069. doi: 10.1371/journal.pone.0006069

[7] Li X W, Yang Y, Henry R J, et al. Plant DNA barcoding: From gene to genome [J]. Biol Rev, 2014, 90(1): 157–166. doi: 10.1111/brv.12104

[8] Li H. The Genus *Paris* (Trilliaceae) [M]. Beijing: Science Press, 1998: 37–194.

[9] Fuse S, Tamura M N. A phylogenetic analysis of the plastid *matK* gene with emphasis on Melanthiaceae *sensu lato* [J]. Plant Biol, 2000, 2(4): 415–427. doi: 10.1055/s-2000-5953

[10] Bremer B, Bremer K, Chase M, et al. An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG III [J]. Bot J Linn Soc, 2009, 161(2): 105–121. doi: 10.1111/j.1095-8339.2009.00996.x

[11] Long C L, Li H, Ouyang Z Q, et al. Strategies for agrobiodiversity conservation and promotion: A case from Yunnan, China [J]. Biodiv Conserv, 2003, 12(6): 1145–1156. doi: 10.1023/A:1023085922265

[12] Zhang T, Liu H, Liu X T, et al. Qualitative and quantitative analysis of steroidal saponins in crude extracts from *Paris polyphylla* var. *yunnanensis* and *P. polyphylla* var. *chinensis* by high performance liquid chromatography coupled with mass spectrometry [J]. J Pharmaceut Biomed Anal, 2010, 51(1): 114–124. doi: 10.1016/j.jpba.2009.08.020

[13] Kang L P, Liu Y X, Eichhorn T, et al. Polyhydroxylated steroidal

glycosides from *Paris polyphylla* [J]. J Nat Prod, 2012, 75(6): 1201–1205. doi: 10.1021/np300045g

[14] Kang L P, Yu K, Zhao Y, et al. Characterization of steroidal glycosides from the extract of *Paris polyphylla* var. *yunnanensis* by UPLC/Q-TOF MSE [J]. J Pharmac Biomed Anal, 2012, 62: 235–249. doi: 10.1016/j.jpba.2011.12.027

[15] Mimaki Y, Kuroda M, Obata Y, et al. Steroidal saponins from the rhizomes of *Paris polyphylla* var. chinensis and their cytotoxic activity on HL-60 cells [J]. Nat Prod Lett, 2000, 14(5): 357–364. doi: 10.1080/10575630008043768

[16] Li Y, Gu J F, Zou X, et al. The anti-lung cancer activities of steroidal saponins of *P. polyphylla* Smith var. *chinensis* (Franch.) Hara through enhanced immunostimulation in experimental Lewis tumor-bearing C57BL/6 mice and induction of apoptosis in the A549 cell line [J]. Molecules, 2013, 18(10): 12916–12936.

[17] Wang G X, Han J, Zhao L W, et al. Anthelmintic activity of steroidal saponins from *Paris polyphylla* [J]. Phytomedicine, 2010, 17(14): 1102–1105. doi: 10.1016/j.phymed.2010.04.012

[18] Do H D K, Kim J S, Kim J H. Comparative genomics of four Liliales families inferred from the complete chloroplast genome sequence of *Veratrum patulum* O. Loes. (Melanthiaceae) [J]. Gene, 2013, 530(2): 229–235. doi: 10.1016/j.gene.2013.07.100

[19] Bodin S S, Kim J S, Kim J H. Complete chloroplast genome of *Chionographis japonica* (Willd.) Maxim. (Melanthiaceae): Comparative genomics and evaluation of universal primers for Liliales [J]. Plant Mol Biol Rep, 2013, 31(6): 1407–1421. doi: 10.1007/s11105-013-0616-x

[20] Do H D K, Kim J S, Kim J H. A *trnI_CAU* triplication event in the complete chloroplast genome of *Paris verticillata* M. Bieb. (Melanthiaceae, Liliales) [J]. Genome Biol Evol, 2014, 6(7): 1699–1706. doi: 10.1093/gbe/evu138

[21] Shi C, Hu N, Huang H, et al. An improved chloroplast DNA extraction procedure for whole plastid genome sequencing [J]. PLoS ONE, 2012, 7(2): e31468. doi: 10.1371/journal.pone.0031468

[22] Dong W P, Xu C, Cheng T, et al. Sequencing angiosperm plastid genomes made easy: A complete set of universal primers and a case study on the phylogeny of Saxifragales [J]. Genome Biol Evol, 2013, 5(5): 989–997.

[23] Haider N. Chloroplast-specific universal primers and their uses in plant studies [J]. Biol Plant, 2011, 55(2): 225–236. doi: 10.1007/s10535-011-0033-7

[24] Stull G W, Moore M J, Mandala V S, et al. A targeted enrichment strategy for massively parallel sequencing of angiosperm plastid genomes [J]. Appl Plant Sci, 2013, 1(2). doi: 10.3732/apps.1200497.

[25] Atherton R A, Mccomish B J, Shepherd L D, et al. Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform [J]. Plant Methods, 2010, 6: 22. doi:10.1186/1746-4811-6-22.

[26] Yang J B, Li D Z, Li H T. Highly effective sequencing whole chloroplast genomes of angiosperms by nine novel universal primer pairs [J]. Mol Ecol Resour, 2014, 14(5): 1024–1031. doi: 10.1111/1755-0998.12251

[27] Patel R K, Jain M. NGS QC Toolkit: A toolkit for quality control of next generation sequencing data [J]. PLoS ONE, 2012, 7(2): e30619. doi: 10.1371/journal.pone.0030619

[28] Matvienko M. CLC Genomics Workbench [CP]. Aarhus: CLC bio, A QIAGEN Company, 2005.

[29] Luo R B, Liu B H, Xie Y L, et al. SOAPdenovo2: An empirically improved memory-efficient short-read *de novo* assembler [J]. GigaScience, 2012, 1: 18. doi:10.1186/2047-217X-1-18

[30] Kearse M, Moir R, Wilson A, et al. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data [J]. Bioinformatics, 2012, 28(12): 1647–1649. doi: 10.1093/bioinformatics/bts199

[31] Echols N, Harrison P, Balasubramanian S, et al. Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes [J]. Nucl Acid Res, 2002, 30(11): 2515–2523.

[32] Tamura K, Stecher G, Peterson D, et al. MEGA6: Molecular evolutionary genetics analysis, Version 6.0 [J]. Mol Biol Evol, 2013, 30(12): 2725–2729. doi: 10.1093/molbev/mst197

[33] Smith S A, Dunn C W. Phyutility: A phyloinformatics tool for trees, alignments and molecular data [J]. Bioinformatics, 2008, 24(5): 715–716. doi: 10.1093/bioinformatics/btm619

[34] Li Q S, Li Y, Song J Y, et al. High-accuracy *de novo* assembly and SNP detection of chloroplast genomes using a SMRT circular consensus sequencing strategy [J]. New Phytol, 2014, 204(4): 1041–1049. doi: 10.1111/nph.12966

[35] Kim J S, Kim J H. Comparative genome analysis and phylogenetic relationship of order Liliales insight from the complete plastid genome sequences of two lilies (*Lilium longiflorum* and *Alstroemeria aurea*) [J]. PLoS ONE, 2013, 8(6): e68180. doi: 10.1371/journal.pone.0068180

[36] Liu J, Qi Z C, Zhao Y P, et al. Complete cpDNA genome sequence of *Smilax china* and phylogenetic placement of Liliales-Influences of gene partitions and taxon sampling [J]. Mol Phylogenet Evol, 2012, 64(3): 545–562. doi: 10.1016/j.ympev.2012.05.010

[37] Kim J S, Kim H T, Yoon C Y, et al. The complete plastid genome sequence of *Bomarea edulis* (Alstroemeriaceae: Liliales) [J/OL]. Mitochondrial DNA, 2014. doi: 10.3109/19401736.2014.971264.

[38] Willey D L, Gray J C. An open reading frame encoding a putative haem-binding polypeptide is contranscribed with the pea chloroplast gene for apocytochrome *f* [J]. Plant Mol Biol,

1990, 15(2): 347–356. doi: 10.1007/BF00036920

[39] Katoh A, Lee K S, Fukuzawa H, et al. *cemA* homologue essential to CO_2 transport in the cyanobacterium *Synechocystis* PCC6803 [J]. Proc Natl Acad Sci USA, 1996, 93(9): 4006–4010.

[40] Logacheva M D, Schelkunov M I, Penin A A. Sequencing and analysis of plastid genome in mycoheterotrophic orchid *Neottia nidus-avis* [J]. Genome Biol Evol, 2011, 3: 1296–1303. doi: 10.1093/gbe/evr102

[41] Logacheva M D, Schelkunov M I, Nuraliev M S, et al. The plastid genome of mycoheterotrophic monocot *Petrosavia stellaris* exhibits both gene losses and multiple rearrangements [J]. Genome Biol Evol, 2014, 6(1): 238–246. doi: 10.1093/gbe/evu001

[42] Goulding S E, Wolfe K H, Olmstead R G, et al. Ebb and flow of the chloroplast inverted repeat [J]. Mol Gen Genet, 1996, 252(1/2): 195–206.

[43] Lin C P, Wu C S, Huang Y Y, et al. The complete chloroplast genome of *Ginkgo biloba* reveals the mechanism of inverted repeat contraction [J]. Genome Biol Evol, 2012, 4(3): 374–381. doi: 10.1093/gbe/evs021

[44] Ji Y H, Fritsch P W, Li H, et al. Phylogeny and classification of *Paris* (Melanthiaceae) inferred from DNA sequence data [J]. Ann Bot, 2006, 98(1): 245–256. doi: 10.1093/aob/mcl095

[45] Moore M J, Soltis P S, Bell C D, et al. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots [J]. Proc Natl Acad Sci USA, 2010, 107(10): 4623–4628. doi: 10.1073/pnas.0907801107

[46] Li X X, Zhou Z K. The higher-level phylogeny of monocots based on *mat*K, *rbc*L and 18S rDNA sequences [J]. Acta Phytotaxon Sin, 2007, 45(2): 113–133.

[47] Zomlefer W B, Williams N H, Whitten W M, et al. Generic circumscription and relationships in the tribe *Melanthieae* (Liliales, Melanthiaceae), with emphasis on *Zigadenus*: Evidence from ITS and *trn*L-F sequence data [J]. Amer J Bot, 2001, 88(9): 1657–1669. doi: 10.2307/3558411.