COMPLETE CHLOROPLAST GENOME OF *EURYA ALATA*, A NECTAR SHRUB THAT BLOSSOMS IN WINTER

CHENG ZHANG[+], YONGFU LI[+], MIN ZHANG[+], YUANYUAN LI, YIFAN DUAN[*] AND XIANRONG WANG[*]

Co-Innovation Center for Sustainable Forestry in Southern China, College of Biology and the Environment,
Nanjing Forestry University, Nanjing, China
[+]*These authors contributed equally.*
[*]*Corresponding author's email: yifan419@hotmail.com; wangxianrong66@njfu.edu.cn.*

Abstract

Eurya alata is one of the few nectar plants that bloom in winter, mainly distributed in the south of the Yangtze River in China. In this research, the chloroplast genome of *E. alata* is assembled and compared with other seven Pentaphylacaceae species. The chloroplast genome of *E. alata* is 157,190 bp and consists of four parts, among which LSC (87,230 bp) and SSC (18,216 bp) are separated by IRa and IRb (51,744 bp). The chloroplast genome encodes 136 genes. They are eight rRNA, 39 tRNA, and 89 protein-coding genes. Besides, 35 SSRs and 49 long-repeat sequences are observed. The protein-coding region of *E. alata* is less variable than the non-coding region. Phylogenetic analysis shows that *Euryodendron excelsum* is the closest species to *E. alata*. In this study, the structure and characteristics of the chloroplast genome of *E. alata* were revealed. These results will be helpful for further research in both *Eurya* and the Pentaphylacaceae family.

Key words: *Eurya alata*; Chloroplast genome; Genomic analysis; Pentaphylacaceae phylogeny.

Introduction

Eurya Thunb. Once the second largest genus of Theaceae, is now subordinated to Pentaphylacaceae (Chase *et al*., 2016). There are approximately 130 species in the genus, mostly located in subtropical and tropical Asia, Hawaiian Islands and other areas of the southwest Pacific. There are greater than 80 species of *Eurya* in China, which is the modern distribution and differentiation center. Besides, *Eurya* plants are an important component of evergreen shrubs in the Yangtze River Basin and southern China. Some plants in this genus can purify the air and absorb heavy metal gas (Pan *et al*., 2006), and extracts from some *Eurya* plants can restrain the merisis of cancer cells (Park *et al*., 2004, 2005).

E. alata is a dioecious evergreen shrub or small tree, with significant scientific research, ecological and economic value. It is also one of the rare precious nectar plants that blossom in winter. The honey is white, transparent and fragrant. It tastes fresh, sweet and is recognized as the king of honey (Pan *et al*., 2006). Meanwhile, tea beverages made by *E. alata* contain tea polyphenols, catechins, and soluble sugar. As a new type of tea drink with low caffeine and high soluble sugar, this tea beverage is of great quality and with an important development value (Wang *et al*., 2016).

As the center of photosynthesis, chloroplast genome contains a great deal of genetic information, which plays a significant part in revealing the mechanism of plant photosynthesis, energy and material metabolism (Zhang & Li, 2011). Shi *et al*., have shown that the transcription mechanism of the chloroplast genome is complicated, that is, the complete transcription occurs not only in the coding regions but also in all its non-coding regions (Shi *et al*., 2016). Moreover, the molecular evolutionary speed of the coding and non-coding regions is significantly different, which can be applied to the systematic study at different levels. More and more researchers are using chloroplast genomes or protein-coding genes of plants to investigate phylogenetic relationships. (Gulden *et al*., 2017; Xiong *et al*., 2018; Xu *et al*., 2020).

Nowadays, there is still little genomic information about the genus *Eurya*, and the chloroplast genomes of most species in Pentaphylacaceae, however, remain unknown. Thus, we sequenced the complete chloroplast genome of *E. alata* and submitted it to GenBank (Accession: MK908406). Then we compared and analyzed the chloroplast genomes of *E. alata* and other species in Pentaphylacaceae, and their phylogenetic relationships were discussed. This study will shed light on the development and utilization of *E. alata* germplasm resources.

Materials and Methods

Genome sequencing and annotation: Fresh leaves of *E. alata* were collected in Xianning, Hubei Province, China. Total DNA was extracted using the improved CTAB method, then sequenced with Illumina HiSeq 2500 platform and utilized NOVOPlasty to assembly the cleaned reads (Doyle & Doyle, 1987; Dierckxsens *et al*., 2017). CpGAVAS was utilized to annotate the genomic structure, including rRNAs, tRNAs, and protein-coding genes (Chang *et al*., 2012). The genome map of *E. alata* was mapped by OGDRAW (Lohse *et al*., 2007). The annotated chloroplast genome was eventually uploaded to GenBank.

Genome analysis and comparison: MEGA7 (Kumar *et al*., 2016) was used for analyzing the relative synonymous codon usage (RSCU) in the chloroplast genome. Long-repeat sequences were detected by the online software REPuter (Kurtz *et al*., 2001), and sequences with different match directions were classified into four categories. Perl script MISA was used to examine mononucleotide and dinucleotide simple sequence repeats (SSRs) (Mudunuri & Nagarajaram, 2007). The chloroplast genome of *E.*

alata, *Euryodendron excelsum*, *Adinandra angustifolia*, *Adinandra millettii*, *Ternstroemia gymnanthera*, *Anneslea fragrans* and *Pentaphylax euryoides* (MK908406, NC_039178, NC_035653, NC_035678, NC_035706, NC_035709, and NC_035710) were compared using Shuffle-LAGAN mode of mVISTA, with *E. alata* as the reference. The boundaries of the junction sites of the chloroplast genomes were visualized by utilizing the online program IRscope (Mayor *et al*., 2000; Amiryousefi *et al*., 2018).

Phylogenetic analysis: Two methods served to construct the phylogenetic relationships of Pentaphylacaceae. On the one hand, MAFFT was initially used to align the whole chloroplast genomes (Nakamura *et al*., 2018), then BioEdit (Hall, 1999) was used to visualize and manually adjust the multiple sequences, GTR was chosen as the optimum base substitution model by jmodelTest2 (Darriba *et al*., 2012), and 1000 bootstrap replicates ML tree was constructed using the RAxML (Stamatakis, 2014). On the other hand, the locally collinear blocks (LCBs) were extracted from chloroplast genomes using HomBlocks (Bi *et al*., 2018), GTR+I+G and GTR +G for different LCBs were selected as the optimum base substitution models using PartitionFinder2 (Lanfear *et al*., 2017), IQ-TREE was used to perform ML tree with 1000 bootstrap replicates (Nguyen *et al*., 2015).

Results and Discussion

Features of *E. alata* chloroplast genome: The cyclic *E. alata* chloroplast genome is 157,190 bp in length, composed of four typical parts, two inverted repeat regions (IRa/IRb; 51,744 bp) are separated by small single-copy region (SSC; 18,216 bp) and large single-copy region (LSC; 87,230 bp) (Fig. 1). The chloroplast genome GC content of *E. alata* is 37.34%. GC content in the four parts is SSC, LSC, and IRa/IRb from low to high, which are 31.04%, 35.31%, and 42.98%, respectively (Table 1). LSC, SSC, and IR regions contain 95, 12 and 29 genes, respectively (Fig. 1). In all, 136 functional genes, including 89 protein-coding genes, eight rRNA genes and 39 tRNA genes were predicted in the *E. alata* chloroplast genome and divided into different groups depending on the gene function (Table 2).

Table 1. Base composition in different parts of the *E. alata* chloroplast genome.

Region	A (%)	C (%)	T (%)	G (%)	GC (%)
LSC	31.68	18.12	33.00	17.20	35.31
SSC	34.39	14.71	34.56	16.33	31.04
IR	28.51	21.49	28.51	21.49	42.98
Total	30.95	18.83	31.71	18.51	37.34

The protein-coding region of *E. alata* chloroplast genome is encoded by 24,003 codons (Table S1), among them, the AUU codon encoding isoleucine appeared the most, with a total of 985, and the UGC codon encoding cysteine appeared the least, with 64 in total. Among all amino acids, leucine and cysteine have the most and the least codons, 2,532 (10.55%) and 268 (1.12%) respectively. In synonymous codons encoding the same amino acid, codons ended with A or U have a higher number and RSCU, codons ended with C or G have a lower number and RSCU. The total GC content of all codons was 38.3%, indicating the preference of AT bases of codons, which situation is also widespread in many other chloroplast genomes (Yi & Kim, 2012; Chen *et al*., 2015; Yu *et al*., 2019).

E. alata chloroplast genome has 16 intron-containing genes, consisting of 7 tRNA and 9 protein-coding genes. Thirteen genes have one intron, while *clpP* and *ycf3* with two introns (Table 3). The *rps12* gene of *E. alata* is a unique trans-splicing gene that contains no introns. Intron deletion of *rps12* also exists in other species, such as *Epipremnum aureum* (Tian *et al*., 2018). As a component of the eukaryotic genome, introns are closely connected with the gene expression process. Introns greatly enrich the number and variety of transcription products and make a complex regulatory role in the splicing process of RNA, which affects gene expression.

Genome differences among diverse species are first manifested by changes in base composition, and GC content plays an important part in genome recognition (Zhu *et al*., 2017). There was poorly difference in GC content among the seven Pentaphylacaceae chloroplast genomes, all of which were about 37%. Besides, *E. alata* has the largest number of genes with 136, followed by *Euryodendron excelsum* with 135 genes (one *ycf1* gene less than *E. alata*). The genes of the other five species were identical, with 132 genes (Table S2). Seven genes lack than *E. alata* are *psbZ*, *rrn5*, *trnN-GUU*, *trnP-GGG*, *trnT-GGU* and two *ycf1* genes; three genes more than *E. alata* are *lhbA*, *rnn5* and *trnG-GCC* (Table S3).

Long-repeat and SSRs analysis: In the *E. alata* chloroplast genome, there were 49 long-repeat sequences identified, including 15 forward repeats (F), 10 reverse repeats (R), 22 palindrome repeats (P) and 2 complement repeats (C). Among them, 10 reverse repeats were 18–23 bp, 15 forward repeats were 18–38 bp, and 22 palindromic repeats were 18–50 bp (Table S4).

Simple sequence repeats (SSRs) are abundant in the entire genome and show high levels of polymorphism. SSRs have consistently been a hotspot in genomic research. They can be dispersed in intron, intergenic, and protein-coding regions. Regions with high genetic diversity also have high mutation rates and polymorphic SSRs. As a novel molecular marker, SSR is widely used in population genetic and phylogenetic analysis, and one of its main sources is chloroplast (Xia *et al*., 2017; Huang *et al*., 2017; Wang *et al*., 2019; Tribhuvan *et al*., 2019). In the *E. alata* chloroplast genome, 35 SSRs were examined, containing 32 mononucleotide SSRs (91.43%) and 3 dinucleotide SSRs (8.57%). SSRs have strong A and T base preferences in composition. Of the 32 mononucleotide SSRs, 12 were A-base repeats, 20 were T-base repeats, and the remaining three dinucleotide SSRs were also consisted of AT base. The longest SSR is multi-base AT repeat with a length of 72 bp (Table 4).

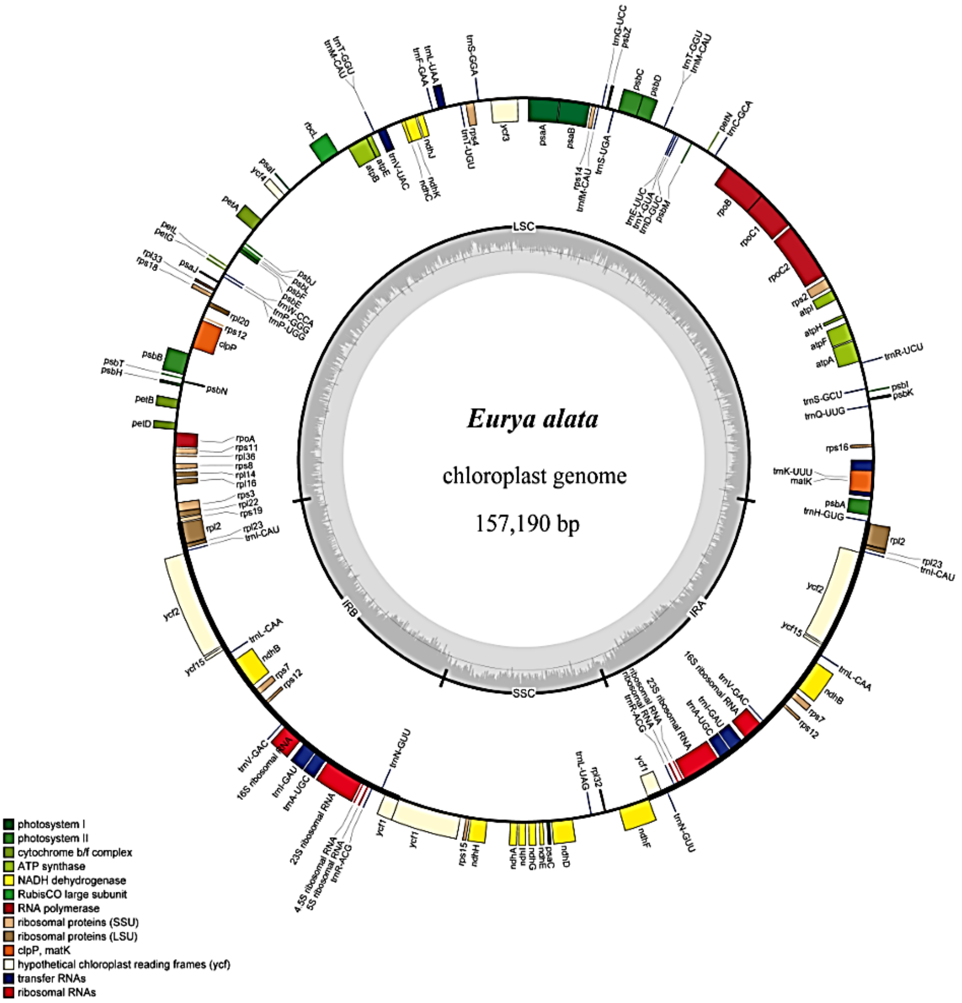

Fig. 1. Gene map of the *Eurya alata* chloroplast genome.

Table 2. Gene annotation of *Eurya alata* chloroplast genome.

Function	Classification	Gene
Self-replication	DNA dependent RNA polymerase	*rpoA*, *rpoB*, *rpoC1**, *rpoC2*
	Large subunit of ribosome	*rpl2*^{2,*}, *rpl14*, *rpl32*, *rpl16*, *rpl20*, *rpl33*, *rpl22*, *rpl23*^{2}, *rpl36*
	Small subunit of ribosome	*rps11*, *rps12*^{2}, *rps14*, *rps2*, *rps16*, *rps3*, *rps18*, *rps4*, *rps7*^{2}, *rps8*, *rps15*, *rps19*
	Transfer RNA genes	*trnA-UGC*^{2,*}, *trnC-GCA*, *trnK-UUU**, *trnD-GUC*, *trnE-UUC*, *trnL-UAG*, *trnF-GAA*, *trnfM-CAU*, *trnG-UCC*, *trnY-GUA*, *trnH-GUG*, *trnI-CAU*^{2}, *trnI-GAU*^{2,*}, *trnL-CAA*^{2}, *trnM-CAU*^{2}, *trnN-GUU*^{2}, *trnP-GGG*, *trnP-UGG*, *trnR-ACG*^{2}, *trnR-UCU*, *trnS-GGA*, *trnS-UGA*, *trnT-GGU*^{2}, *trnT-UGU*, *trnV-GAC*^{2}, *trnS-GCU*, *trnV-UAC**, *trnW-CCA*, *trnQ-UUG*, *trnL-UAA**,
	Ribosomal RNA genes	*rrn5*^{2}, *rrn16*^{2}, *rrn23*^{2}, *rrn4.5*^{2},
Photosynthesis	ATP synthase	*atpA*, *atpF**, *atpI*, *atpE*, *atpB*, *atpH*
	Photosystem I	*psaB*, *psaA*, *psaI*, *psaC*, *psaJ*
	Photosystem II	*psbA*, *psbK*, *psbI*, *psbM*, *psbD*, *psbB*, *psbC*, *psbZ*, *psbE*, *psbF*, *psbT*, *psbH*, *psbJ*, *psbL*, *psbN*
	Cytochrome b/f complex	*petL*, *petB*, *petD*, *petG*, *petA*, *petN*
	Large subunit of rubisco	*rbcL*
	NADH dehydrogenase	*ndhE*, *ndhA*, *ndhJ*, *ndhB*^{2,*}, *ndhC*, *ndhF*, *ndhG*, *ndhI*, *ndhK*, *ndhD*, *ndhH*
Other genes	Translational initiaton factor	*infA*
	ATP-dependent protease subunit gene	*clpP***
	Subunit of acetyl-CoA-carboxylase	*accD*
	C-type cytochrome synthesis gene	*ccsA*
	Envelope membrane protein	*cemA*
	Maturase	*matK*
	Unknow function	*ycf1*^{3}, *ycf2*^{2}, *ycf3***, *ycf4*, *ycf15*^{2}

* Number of introns; ^{2,3} Copy number of genes

Table 3. Exons and introns size of genes with introns in the *E. alata* chloroplast genome.

Gene	Distribution	Exon I (bp)	Intron I (bp)	Exon II (bp)	Intron II (bp)	Exon III (bp)
atpF	LSC	411	710	159		
clpP	LSC	219	668	291	822	69
ndhB	IRa	756	679	777		
ndhB	IRb	777	679	756		
rpl2	IRa	435	662	393		
rpl2	IRb	393	662	435		
rpoC1	LSC	1626	731	456		
rps12	LSC	114	-			
rps12	IRa	240	-			
rps12	IRb	240	-			
trnA-UGC	IRa	38	807	35		
trnA-UGC	IRb	35	807	38		
trnI-GAU	IRa	42	944	35		
trnI-GAU	IRb	35	944	42		
trnK-UUU	LSC	35	2526	37		
trnL-UAA	LSC	37	508	50		
trnV-UAC	LSC	37	586	39		
ycf3	LSC	153	737	228	713	126

Table 4. SSRs examined in the *E. alata* chloroplast genome.

Type	SSR	Length	Start	End	Type	SSR	Length	Start	End
p1	(A)11	11	3348	3358	p1	(T)10	10	61239	61248
p1	(A)11	11	5626	5636	p1	(T)11	11	62024	62034
p1	(A)10	10	6553	6562	p1	(T)11	11	63663	63673
p1	(T)13	13	6858	6870	p1	(T)11	11	66230	66240
p1	(T)10	10	7468	7477	p1	(T)10	10	71437	71446
p1	(A)10	10	7898	7907	p1	(T)11	11	73549	73559
p1	(T)10	10	8970	8979	c	(A)10(A)11	22	73710	73731
p1	(T)10	10	11190	11199	p1	(A)12	12	74345	74356
p1	(A)13	13	13635	13647	p1	(T)10	10	81199	81208
p1	(A)11	11	17509	17519	p1	(T)13	13	83595	83607
p1	(T)11	11	19720	19730	p1	(T)12	12	85652	85663
p1	(T)10	10	27419	27428	p1	(A)10	10	110803	110812
p1	(T)10	10	33384	33393	p1	(A)11	11	116014	116024
p1	(A)11	11	43845	43855	c	(A)13(T)11	72	125718	125789
p1	(A)12	12	48889	48900	c	(A)10(A)10	27	126922	126948
p1	(T)12	12	53062	53073	p1	(A)11	11	128382	128392
p1	(T)10	10	56781	56790	p1	(T)10	10	133609	133618
p1	(T)11	11	59446	59456					

Comparative chloroplast genomic analysis of seven Pentaphylacaceae species: Although chloroplast genomes are conservative among related species, there are still some differences. Chloroplast genomes of seven species in Pentaphylacaceae including *E. alata*, *Adinandra angustifolia*, *Adinandra millettii*, *Anneslea fragrans, Pentaphylax euryoides*, *Ternstroemia gymnanthera*, and *Euryodendron excelsum* were compared using mVISTA, and *E. alata* was set as the reference genome (Fig. 2). IRa and IRb regions possessed higher consistency than SSC and LSC, which is also related to the more conservative of IR regions in the evolutionary process. In conserved non-coding sequences (CNS), 4-10k, 28-34k, 53-54k, 114-118k and other regions have considerable divergence. On the contrary, the exon regions and the untranslated regions (UTRs) have small divergence. Generally speaking, the protein-coding region showed strong conservation, and its consistency was higher than the non-coding region.

Boundary analysis of four regions: The chloroplast genome is a circular structure with four boundaries between IRa/IRb, LSC and SSC, that is, LSC/IRB (JLB), LSC/IRS (JLA), SSC/IRB (JSB) and SSC/IRA (JSA) (Fig. 3). The contraction and expansion of IR boundaries during genome evolution often lead to some genes entering IR region or single-copy region and may reflect the evolutionary relationships between species (Wang *et al*., 2017). The comparison results of the IR region boundaries show that *E. alata* and *Euryodendron excelsum* were closest in the boundary structure and gene order. The *ndhf* coding regions of these two species are at the JSA boundary, and the *ycf1* coding region is at the JSB boundary. Moreover, the order of genes in the SSC region is exactly the opposite, so we speculate it was caused by the reversal of the SSC regions of these two species, and specific reasons require further research. At the JLB is the *rps19* coding region, however, the *rpl22* coding region of *Anneslea fragrans* enters this boundary, resulting in the distribution of the *rps19* coding region in the IRb region. In addition, the chloroplast genome of *Anneslea fragrans* has only one tiny *rpl2* coding region, while the others contain two. The *trnH* coding region of the seven species is downstream of the JLA and the distance from the boundary is 1-22 bp.

Phylogenetic Analysis

We constructed two maximum likelihood (ML) phylogenetic trees of 14 species with whole chloroplast genomes and locally collinear blocks (LCBs) extracted from 14 chloroplast genomes (Fig. S1), respectively. Both trees had identical phylogenetic topologies and most of the branches have high bootstrap support (Fig. 4). *E. alata* is far related to *Camellia japonica* and other species of Theaceae and is clustered with Pentaphylacaceae species. *E. alata* and *Euryodendron excelsum* gather in the same branch and form a sister relationship with *Adinandra* (tribe Freziereae). Within all the Pentaphylacaceae species, *Pentaphylax euryoidesare* (tribe Pentaphylaceae) is the earliest branch followed by *Ternstroemia gymnanthera* and *Anneslea fragrans* clade (tribe Ternstroemieae), this phylogenetic topology is consistent with APGIV (Chase *et al*., 2016) and previous research (Shi *et al*., 2018).

It is worth mentioning that the monophyletic clade formed by *Sladenia celastrifolia* and Pentaphylacaceae species with 100% bootstrap support in both two phylogenetic trees. Previous results based on DNA sequences also suggested that *Sladenia* and Pentaphylacaceae are very close and proposed to merge them into a single family according to other morphological and embryological characteristics (Savolainen *et al*., 2000; Yu *et al*., 2017; Tsou *et al*., 2016; Rose *et al*., 2018). The phylogenetic topologies of two ML trees constructed by chloroplast genomes and LCBs respectively support the previous results well.

Supplementary Materials

Table S1. Relative synonymous codon usage (RSCU) for protein-coding genes of *E. alata* chloroplast genome.

Amino acid	Codon	Count	RSCU	tRNA	Amino acid	Codon	Count	RSCU	tRNA
Phe	UUU(F)	858	1.27		Tyr	UAU(Y)	722	1.64	
Phe	UUC(F)	493	0.73	*trnF-GAA*	Tyr	UAC(Y)	160	0.36	*trnY-GUA*
Leu	UUA(L)	792	1.88		Stop	UAA(*)	41	1.45	
Leu	UUG(L)	538	1.27	*trnL-CAA*	Stop	UAG(*)	23	0.81	
Leu	CUU(L)	537	1.27		His	CAU(H)	462	1.57	
Leu	CUC(L)	175	0.41		His	CAC(H)	125	0.43	*trnH-GUG*
Leu	CUA(L)	332	0.79		Gln	CAA(Q)	637	1.51	*trnQ-UUG*
Leu	CUG(L)	158	0.37		Gln	CAG(Q)	207	0.49	
Ile	AUU(I)	985	1.44		Asn	AAU(N)	837	1.51	
Ile	AUC(I)	420	0.61	*trnI-CAU*	Asn	AAC(N)	274	0.49	
Ile	AUA(I)	649	0.95		Lys	AAA(K)	876	1.48	
Met	AUG(M)	583	1	*trnM-CAU*	Lys	AAG(K)	306	0.52	
Val	GUU(V)	491	1.48		Asp	GAU(D)	794	1.62	
Val	GUC(V)	155	0.47	*trnV-GAC*	Asp	GAC(D)	187	0.38	*trnD-GUC*
Val	GUA(V)	487	1.47		Glu	GAA(E)	904	1.49	*trnE-UUC*
Val	GUG(V)	190	0.57		Glu	GAG(E)	307	0.51	
Ser	UCU(S)	550	1.79		Cys	UGU(C)	204	1.52	
Ser	UCC(S)	299	0.97	*trnS-GGA*	Cys	UGC(C)	64	0.48	*trnC-GCA*
Ser	UCA(S)	360	1.17	*trnS-UGA*	Stop	UGA(*)	21	0.74	
Ser	UCG(S)	178	0.58		Trp	UGG(W)	432	1	*trnW-CCA*
Pro	CCU(P)	408	1.6		Arg	CGU(R)	342	1.39	*trnR-ACG*
Pro	CCC(P)	189	0.74		Arg	CGC(R)	76	0.31	
Pro	CCA(P)	299	1.17	*trnP-UGG*	Arg	CGA(R)	353	1.44	
Pro	CCG(P)	124	0.49		Arg	CGG(R)	114	0.46	
Thr	ACU(T)	501	1.64		Ser	AGU(S)	348	1.13	
Thr	ACC(T)	231	0.76	*trnT-GGU*	Ser	AGC(S)	106	0.35	*trnS-GCU*
Thr	ACA(T)	365	1.19	*trnT-UGU*	Arg	AGA(R)	424	1.73	*trnR-UCU*
Thr	ACG(T)	126	0.41		Arg	AGG(R)	162	0.66	
Ala	GCU(A)	619	1.85		Gly	GGU(G)	550	1.31	
Ala	GCC(A)	208	0.62		Gly	GGC(G)	175	0.42	*trnG-GCC*
Ala	GCA(A)	376	1.12		Gly	GGA(G)	691	1.64	
Ala	GCG(A)	137	0.41		Gly	GGG(G)	266	0.63	

RSCU: Relative Synonymous Codon Usage

Table S2. Genome features of seven Pentaphylacaceae species.

	Eurya alata	*Euryodendron excelsum*	*Adinandra angustifolia*	*Adinandra millettii*	*Ternstroemia gymnanthera*	*Anneslea fragrans*	*Pentaphylax euryoides*
Accession	MK908406	NC_039178	NC_035653	NC_035678	NC_035706	NC_035709	NC_035710
Total	157190	157702	156344	156311	157753	158182	156132
LSC	87230	87144	85743	85698	87225	86524	85635
SSC	18216	18414	18419	18421	18410	18398	18341
IR	51744	52144	52182	52192	52118	53260	52156
Number of genes	136	135	132	132	132	132	132
protein–coding genes	89	88	87	87	87	87	87
tRNA genes	39	39	37	37	37	37	37
rRNA genes	8	8	8	8	8	8	8
GC content (%)	37.3	37.3	37.4	37.4	37.2	37.2	37.1

Table S3. Differential genes in seven Pentaphylacaceae chloroplast genomes.

Species	*lhbA*	*psbZ*	*rnn5*	*rrn5*	*trnG-GCC*	*trnN-GUU*	*trnP-GGG*	*trnT-GGU*	*ycf1*
Eurya alata	0	1	0	1	0	1	1	1	3
Euryodendron excelsum	0	1	0	1	0	1	1	1	2
Adinandra angustifolia	1	0	1	0	1	0	0	0	1
Adinandra millettii	1	0	1	0	1	0	0	0	1
Anneslea fragrans	1	0	1	0	1	0	0	0	1
Pentaphylax euryoides	1	0	1	0	1	0	0	0	1
Ternstroemia gymnanthera	1	0	1	0	1	0	0	0	1

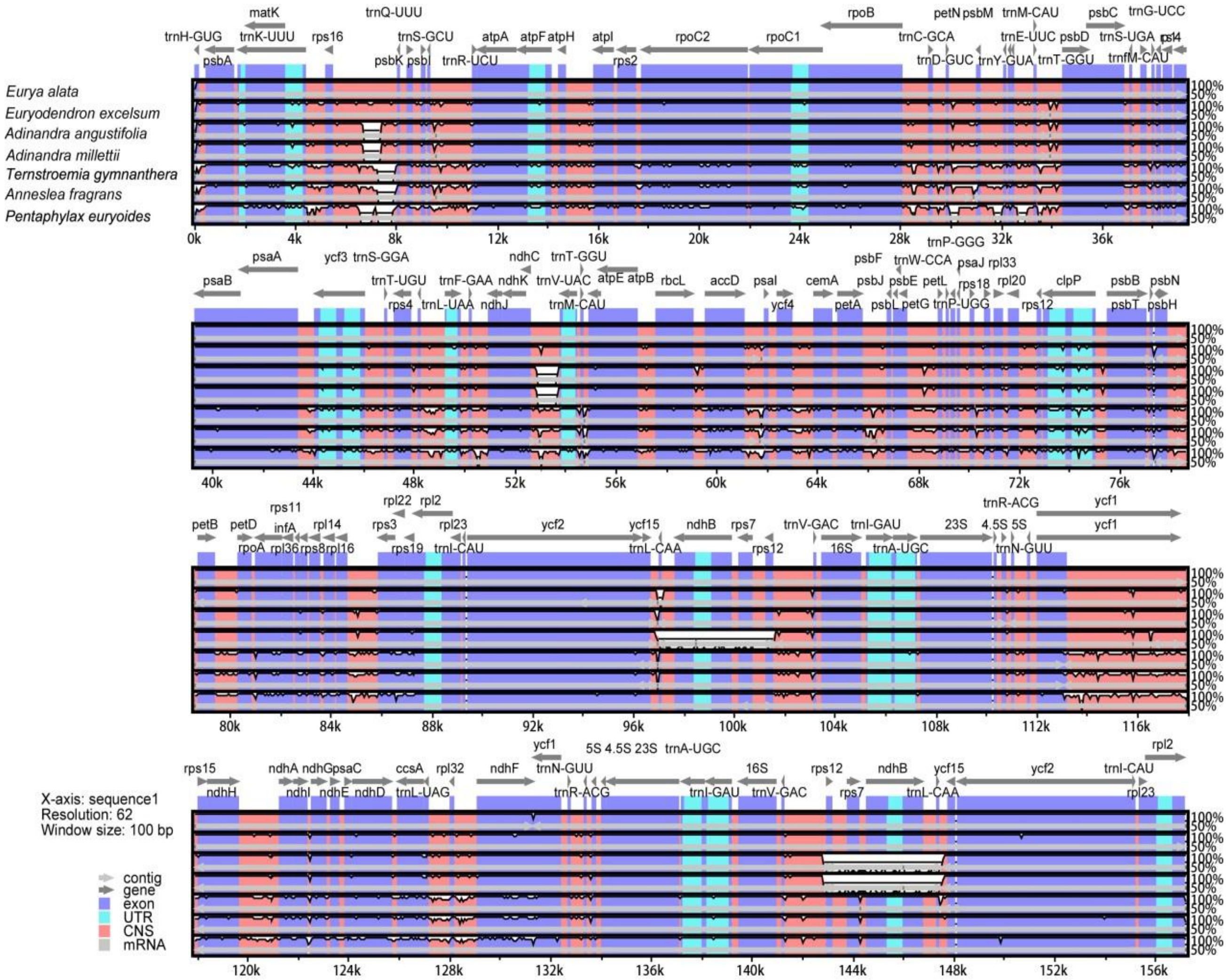

Fig. 2. Genome comparison of seven Pentaphylacaceae chloroplast genomes. The dark blue, light blue and pink regions represent exons, UTRs and CNS, respectively. The y-axis represents the percent of conservation.

Table S4. Identification of long-repeat sequences in *E. alata* chloroplast genome.

Size (bp)	Starting position I	Match direction II	Starting position	Distance of repeat	E-value	Location
50	77318	P	77318	0	5.48E-21	LSC
38	94451	F	94487	0	9.20E-14	IRb
38	94451	P	149895	0	9.20E-14	IRb;IRa
38	94487	P	149931	0	9.20E-14	IRb;IRa
38	149895	F	149931	0	9.20E-14	IRa
36	79790	P	79790	0	1.47E-12	LSC
35	94459	F	94477	-1	6.18E-10	IRb
35	94459	P	149908	-1	6.18E-10	IRb;IRa
35	94477	P	149926	-1	6.18E-10	IRb;IRa
35	149908	F	149926	-1	6.18E-10	IRa
30	9242	P	46890	0	6.03E-09	LSC
31	61682	P	61682	-1	1.40E-07	LSC
27	10809	P	10845	0	3.86E-07	LSC
27	30043	F	30068	0	3.86E-07	LSC
30	94477	F	94495	-1	5.42E-07	IRb
30	94477	P	149895	-1	5.42E-07	IRb;IRa
30	94495	P	149913	-1	5.42E-07	IRb;IRa
30	149895	F	149913	-1	5.42E-07	IRa
26	90811	P	90811	0	1.54E-06	IRb
26	90811	F	153583	0	1.54E-06	IRb;IRa
26	153583	P	153583	0	1.54E-06	IRa
28	33463	P	33469	-1	8.10E-06	LSC
28	101552	P	121114	-1	8.10E-06	IRb
28	121114	F	142840	-1	8.10E-06	IRa
24	112647	P	112647	0	2.47E-05	IRb
24	112647	F	131749	0	2.47E-05	IRb;IRa
24	131749	P	131749	0	2.47E-05	IRa
27	45222	F	101552	-1	3.12E-05	LSC;IRb
27	45222	P	142841	-1	3.12E-05	LSC;IRa
23	30855	P	30882	0	9.88E-05	LSC
23	48751	R	48751	0	9.88E-05	LSC
22	32496	P	32496	0	3.95E-04	LSC
22	38130	P	38130	0	3.95E-04	LSC
22	96927	P	96953	0	3.95E-04	IRb
22	96927	F	147445	0	3.95E-04	IRb;IRa
22	96953	F	147471	0	3.95E-04	IRb;IRa
22	101558	P	121114	0	3.95E-04	IRb
22	147445	P	147471	0	3.95E-04	IRa
25	82409	F	82433	-1	4.63E-04	LSC
25	94451	F	94469	-1	4.63E-04	IRb
25	94451	P	149926	-1	4.63E-04	IRb;IRa
25	94469	P	149944	-1	4.63E-04	IRb;IRa
25	149926	F	149944	-1	4.63E-04	IRa
21	9248	F	37069	0	1.58E-03	LSC
21	28400	F	28420	0	1.58E-03	LSC
21	37069	P	46893	0	1.58E-03	LSC
21	38190	F	69566	0	1.58E-03	LSC
21	73709	R	73709	0	1.58E-03	LSC
24	10745	F	38001	-1	1.78E-03	LSC

F, Forward repeat; R, Reverse repeat; C, Complement repeat; P, Palindrome repeat

Supplementary materials: Table S1. Relative synonymous codon usage (RSCU) for protein-coding genes of *E. alata* chloroplast genome. Table S2. Genome features of seven Pentaphylacaceae species. Table S3. Differential genes in seven Pentaphylacaceae chloroplast genomes. Table S4. Identification of long-repeat sequences in *E. alata* chloroplast genome. Fig. S1. Locally collinear blocks of 14 chloroplast genomes generated by HomBlocks. The total length of locally collinear blocks is 87,149 bp, the black parts and gray parts represent the bases variation locus and consistent locus, respectively.

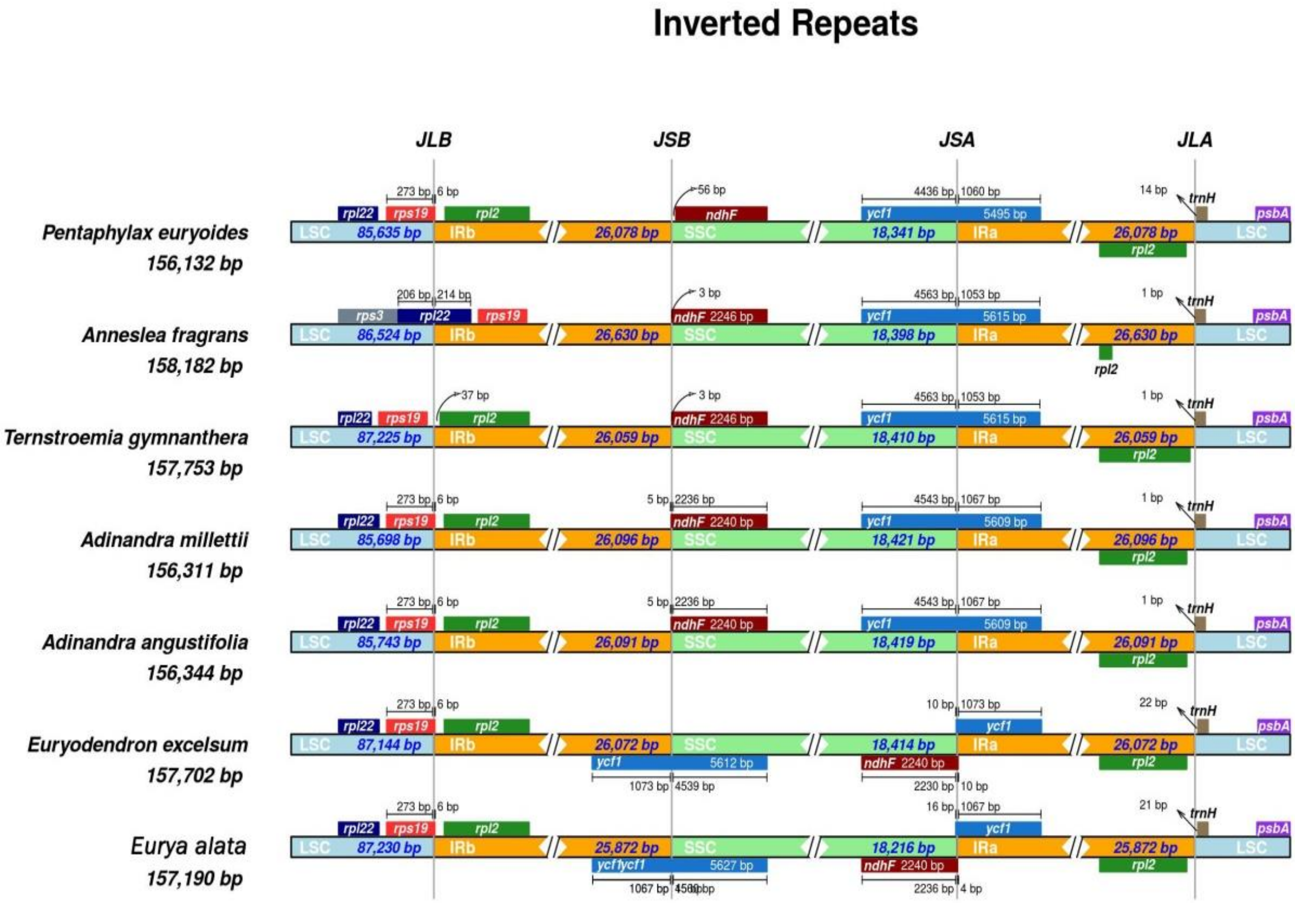

Fig. 3. Boundary analysis of four regions among 7 Pentaphylacaceae species.

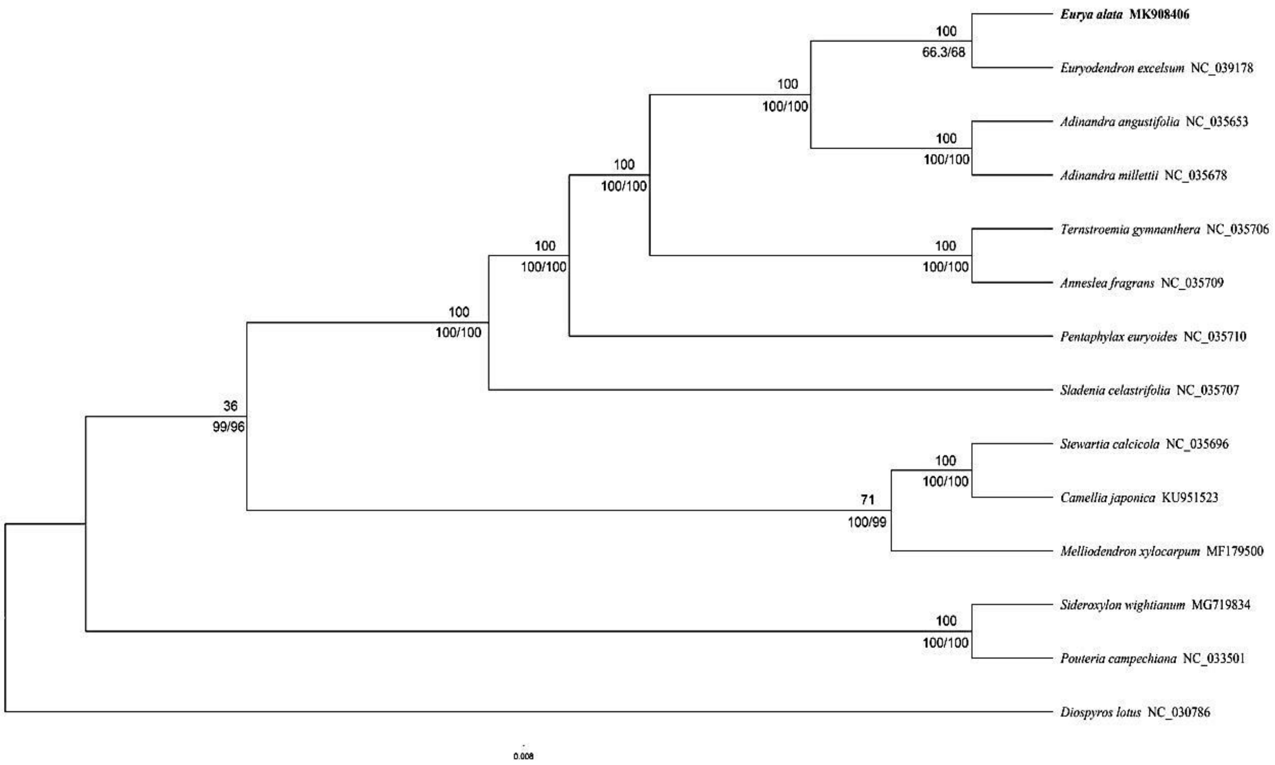
Eurya alata MK908406
Euryodendron excelsum NC_039178
Adinandra angustifolia NC_035653
Adinandra millettii NC_035678
Ternstroemia gymnanthera NC_035706
Anneslea fragrans NC_035709
Pentaphylax euryoides NC_035710
Sladenia celastrifolia NC_035707
Stewartia calcicola NC_035696
Camellia japonica KU951523
Melliodendron xylocarpum MF179500
Sideroxylon wightianum MG719834
Pouteria campechiana NC_033501
Diospyros lotus NC_030786
100
66.3/68
100
100/100
100
100/100
100
100/100
100
100/100
100
100/100
100
100/100
36
99/96
100
100/100
71
100/99
100
100/100
0.008

Fig. 4. Phylogenetic tree of 14 species constructed by whole chloroplast genomes and locally collinear blocks (LCBs). The bootstrap supports of the phylogenetic tree constructed by chloroplast genomes are above the branches, and numbers below the branches indicate the bootstrap support (SH-aLRT support / ultrafast bootstrap support) of the phylogenetic tree constructed by locally collinear blocks (LCBs) extracted from chloroplast genomes.

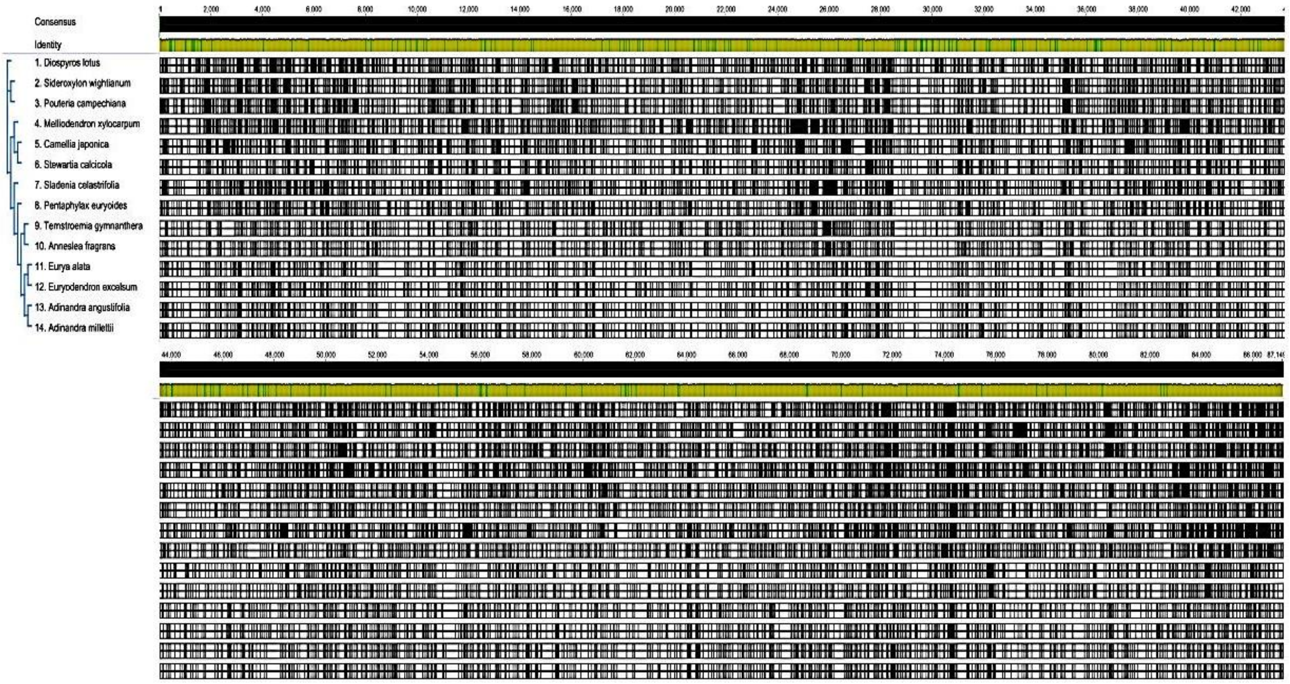

Fig. S1. Locally collinear blocks of 14 chloroplast genomes generated by HomBlocks. The total length of locally collinear blocks is 87,149 bp, the black parts and gray parts represent the bases variation locus and consistent locus, respectively.

Conclusions

The noncoding region of *E. alata* chloroplast genome has more mutation hotspots and faster mutation rate than the protein-coding region, which can be employed in population genetics research. The results of the genome structure comparison suggested that inversion occurred in the SSC region of *E. alata* and *Euryodendron excelsum*, the specific reasons are not clear and further research is needed. In phylogenetic analysis, some branches support of phylogenetic tree constructed based on whole chloroplast genome sequence is low, thus, we prefer to construct phylogenetic trees with locally collinear blocks shared by the chloroplast genomes, especially in the case of genome rearrangements. At present, only a small proportion of plants chloroplast genomes have been sequenced in the genus *Eurya* and Pentaphylacaceae family, and the unresolved issues depend on the publication of more chloroplast genome sequences in the future.

Acknowledgements

This work was supported by Natural Science Foundation of Jiangsu Province (BK20160932), the Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX19_1073), China Postdoctoral Science Foundation (2016M590462), and the Priority Academic Program Development (PAPD) of Jiangsu High Education Institutions, Jiangsu Province, China. The authors thank Min Tao and Shujie Xiang from Nanjing Forestry University for their assistance with data analyses.

References

Amiryousefi, A., J. Hyvönen and P. Poczai. 2018. IRscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics*, 34(17): 3030-3031.

Bi, G., Y. Mao, Q. Xing and M. Cao. 2018. HomBlocks: a multiple-alignment construction pipeline for organelle phylogenomics based on locally collinear block searching. *Genom.*, 110(1): 18-22.

Chang, L., L. Shi, Y. Zhu, H. Chen, J. Zhang, X. Lin and X. Guan. 2012. CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. *BMC Genom.*, 13(1):715.

Chase, M., M. Christenhusz, M. Fay, J. Byng, W. Judd, D. Soltis, D. Mabberley, A. Sennikov and P. Soltis. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot. J. Linn. Soc.*, 181(1):1-20.

Chen, J., Z. Hao, H. Xu, L. Yang, G. Liu, Y. Sheng, C. Zheng, W. Zheng, T. Cheng and J. Shi. 2015. The complete chloroplast genome sequence of the relict woody plant *Metasequoia glyptostroboides* Hu et Cheng. *Front. Plant Sci.*, 6: 447.

Darriba, D., G. Taboada, R. Doallo and D. Posoda. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods*, 9(8): 772-772.

Dierckxsens, N., P. Mardulyn and G. Smits. 2017. Novoplasty: De novo assembly of organelle genomes from whole genome DNA. *Nucl. Acids Res.*, 45(4): e18.

Doyle, J. and J. Doyle. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.*, 19: 11-15.

Gulden, D., Y. Asli, B. Eyup and K. Zekİ. 2017. Molecular phylogeny of section *Drosanthe* (Spach) Endl. (*Hypericum* L.) inferred from chloroplast genome. *Pak. J. Bot.*, 49(6): 2235-2242.

Hall, T.A. 1999. BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids Symp. Ser.*, 41(41): 95-98.

Huang, J., R. Chen and X. Li. 2017. Comparative analysis of the complete chloroplast genome of four known *Ziziphus* species. *Genes*, 8(12): 340.

Kumar, S., G. Stecher and K. Tamura. 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.*, 33(7): 1870-1874.

Kurtz, S., J. Choudhuri, E. Ohlebusch, C. Schleiermacher, J. Stoye and R. Giegerich. 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucl. Acids Res.*, 29(22): 4633-4642.

Lanfear, R., P. Frandsen, A. Wright, T. Senfeld and B. Calcott. 2017. Partition Finder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Mol. Biol. Evol.*, 34(3): 772-773.

Lohse, M., O. Drechsel and R. Bock. 2007. Organellar Genome DRAW(OGDRAW): A tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Curr. Genet.*, 52(5-6): 267-274.

Mayor, C., M. Brudno, J. Schwartz, A. Poliakov, E. Rubin, K. Frazer, L. Pachter and I. Dubchak. 2000. VISTA: visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*, 16(11): 1046-1047.

Mudunuri, S. and H. Nagarajaram. 2007. IMEx: imperfect microsatellite extractor. *Bioinformatics*, 23(10): 1181-1187.

Nakamura, T., K. Yamada, K. Tomii and K. Katoh. 2018. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics*, 34(14): 2490-2492.

Nguyen, L., H. Schmidt, A. Haeseler and B. Minh. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating Maximum-Likelihood phylogenies. *Mol. Biol. Evol.*, 32(1): 268-274.

Pan, J., D. Ji and G. Tang. 2006. The wild plant resources and application of *Eurya* in China. *Chin. Wild Plant Resour.*, 25(2): 36-38 (in Chinese).

Park, S., H. Lee, W. Yoon, G. Kang, J. Moon, N. Lee, S. Kim, H. Kang and E. Yoo. 2005. Inhibitory effects of eutigosides isolated from *Eurya emarginataon* the inflammatory mediators in RAW264.7 cells. *Arch. Pharm. Res.*, 28(11): 1244-1250.

Park, S., H. Yang, J. Moon, N. Lee, S. Kim, J. Kang, Y. Lee, D. Park, E. Yoo and H.K. Kang. 2004. Induction of the apoptosis of HL-60 promyelocytic leukemia cells by *Eurya emarginata*. *Cancer Lett.*, 205(1): 31-38.

Rose, J., T. Kleist, S. Löfstrand, B. Drew, J. Schönenberger and K. Sytsma. 2018. Phylogeny, historical biogeography, and diversification of angiosperm order Ericales suggest ancient Neotropical and East Asian connections. *Mol. Phylogen. Evol.*, 122: 59-79.

Savolainen, V., M. Fay, D. Albach, A. Backlund, M. Bank, K. Cameron, S. Johnson, M. Lledó, J. Pintaud, M. Powell, M. Sheahan, D. Soltis, P. Soltis, P. Weston, W. Whitten, K. Wurdack and M. Chase. 2000. Phylogeny of the Eudicots: a nearly complete familial analysis based on rbcL gene sequences. *Kew Bull.*, 55: 257-309.

Shi, C., S. Wang, E. Xia, J. Jiang, F. Zeng and L. Gao. 2016. Full transcription of the chloroplast genome in photosynthetic eukaryotes. *Sci. Rep.*, 6: 30135.

Shi, X., W. Li, F. Xing, Y. Zhou, W. Guo and Y. Huang. 2018. Characterization of the complete chloroplast genome of *Euryodendron excelsum* (Pentaphylacaceae), a critically endangered species endemic to China. *Conserv. Genet. Resour.*, 11(3): 275-278.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9): 1312-1313.

Tian, N., L. Han, C. Chen and Z. Wang. 2018. The complete chloroplast genome sequence of *Epipremnum aureum* and its comparative analysis among eight *Araceae* species. *PloS One*, 13(3): e0192956.

Tribhuvan, K., A. Mithra, P. Sharma, A. Das, K. Kumar, A. Tyagi, A. Solanke, Sandhya, R. Sharma, P. Jadhav, M. Raveendran, B. Fakrudin, T. Sharma, N. Singh and K. Gaikward. 2019. Identification of genomic SSRs in cluster bean (*Cyamopsis tetragonoloba*) and demonstration of their utility in genetic diversity analysis. *Ind. Crop. Prod.*, 133: 221-231.

Tsou, C., L. Li and K. Vijayan. 2016. The intra-familial relationships of Pentaphylacaceae s.l. as revealed by DNA sequence analysis. *Biochem. Genet.*, 54(3): 270-282.

Wang, F., S. Zhao, H. Shen, L. Zhao and T. Yan. 2016. Study on biological and ecological characteristics, resource situation, protection and exploitation of ancient populations of *Eurya alata* in Dachagou valley of Xinyang. *J. Henan Agri. Sci.*, 45(8): 43-48 (in Chinese).

Wang, Y., Z. Zeng, F. Li, X. Yang, X. Gao, Y. Ma, J. Rao, H. Wang and T. Liu. 2019. A genomic resource derived from the integration of genome sequences, expressed transcripts and genetic markers in ramie. *BMC Genom.*, 20(1): 476.

Wang, Y., X. Qu, S. Chen, D. Li and T. Yi. 2017. Plastomes of Mimosoideae: structural and size variation, sequence divergence, and phylogenetic implication. *Tree Genet. Genom.*, 13(2): 41.

Xia, X., L. Luan, G. Qin, L. Yu, Z. Wang, W. Dong, Y. Song, Y. Qiao, X. Zhang, Y. Sang and L. Yang. 2017. Genome-wide analysis of SSR and ILP markers in trees: diversity profiling, alternate distribution, and applications in duplication. *Sci. Rep.*, 7(1): 17902.

Xiong, B., L. Zhang, L. Xie, S. Dong and Z. Zhang. 2018. Complete chloroplast genome of a valuable economic tree, *Lindera glauca* (Lauraceae) and comparison with its congeners. *Pak. J. Bot.*, 50(6): 2189-2196.

Xu, Z., D. Yuan, Y. Tang, L. Wu and Y. Zhao. 2020. *Camellia hainanica* (Theaceae) a new species from Hainan, supported from morphological characters and phylogenetic analysis. *Pak. J. Bot.*, 52(3): 1025-1032.

Yi, D. and K. Kim. 2012. Complete chloroplast genome sequences of important oilseed crop *Sesamum indicum* L. *PloS One*, 7(5): e35872.

Yu, X., L. Gao, D. Soltis, P. Soltis, J. Yang, L. Fang, S. Yang and D. Li. 2017. Insights into the historical assembly of East Asian subtropical evergreen broadleaved forests revealed by the temporal history of the tea family. *New Phytol.*, 215(3): 1235-1248.

Yu, X., L. Zuo, D. Lu, B. Lu, M. Yang and J. Wang. 2019. Comparative analysis of chloroplast genomes of five *Robinia* species: genome comparative and evolution analysis. *Gene*, 689: 141-151.

Zhang, Y. and D. Li. 2011. Advances in phylogenomics based on complete chloroplast genomes. *Plant Divers. Resour.*, 33(4): 365-375 (in Chinese).

Zhu, T., L. Zhang, W. Chen, J. Yin and Q. Li. 2017. Analysis of chloroplast genomes in 1,342 plants. *Genom. Appl. Biol.*, 36(10): 4323-4333 (in Chinese).

(Received for publication 16 August 2019)