Journal of Genetics (2020) 99:85
https://doi.org/10.1007/s12041-020-01243-5

RESEARCH ARTICLE

Plastid genomes of *Elaeagnus mollis*: comparative and phylogenetic analyses

YONGQIN CHENG[1], YANCI YANG[2], XUENING FU[1], LI LIU[1], ZAIMIN JIANG[3] and JING CAI[1]*

[1]*College of Forestry, Northwest A&F University, Yangling 712100, People's Republic of China*
[2]*School of Biological Science and Technology, Baotou Teachers' College, Baotou 014030, People's Republic of China*
[3]*College of Life Science, Northwest A&F University, Yangling 712100, People's Republic of China*
*For correspondence. E-mail: cjcaijing@163.com.

Received 16 November 2019; revised 11 June 2020; accepted 13 August 2020; published online 11 November 2020

Abstract. Plastomes, which are maternally inherited and show a moderate rate of evolution, play a critical role in phylogenetic reconstruction and assignment of plant species. However, little is known about the sequence divergence and molecular evolutionary patterns of plastid genomes in *Elaeagnus mollis*, a plant of great economic, medicinal, edible and ecological values. The plastid genome of *E. mollis* is 152,224-bp long and has 47 repeat sequences, including tandem (17), dispersed (12), and palindromic (18) types of repeat variations. Here, we reported six divergence hotspots (*atpH-atpI*, *petN-psbM*, *trnT-psbD*, *trnP-psaJ*, *rpl32-trnL* and *ycf1*) that could potentially be used as molecular genetic markers for population genetics and phylogenetic studies of *E. mollis*. A comparison of plastid genomes in the order Rosales showed that the *trnH* gene was duplicated only in Elaeagnaceae; therefore, it is an important marker in Elaeagnaceae. Phylogenetic analyses based on whole plastid genome sequences in 33 species revealed that Rosales is divided into two strongly supported clades and that the families Elaeagnaceae and Barbeyaceae are closely related.

Keywords. plastid genome; phylogenetic relationship; repeat sequence; *trnH* gene; *Elaeagnus mollis*.

Introduction

Elaeagnus mollis is known as 'monk Tang's flesh' because of its seed, which contains vitamin E up to 1558.1 mg/100 g (Yao 2005), which is rare in nature. *E. mollis*, a small deciduous tree belongs to the genus *Elaeagnus* of the family Elaeagnaceae, is a relic of Quaternary glaciations in China. It is regarded as a rare woody oil plant with high economic, medicinal, edible and ecological values (Xie and Ling 1997; Liang *et al.* 2015). Previous studies showed that *E. mollis* kernels have 32.2l% protein, with 17 different amino acids. Seven of these amino acids (valine, methionine, leucine, isoleucine, phenylalanine and lysine) are essential for humans and animals (Yao 2005). Copious high-quality grease, especially linoleic acid is found in *E. mollis* seeds, which can be used to treat arteriosclerosis (Zhang and Zhang 2015). However, such precious resources are only distributed in the Shaanxi and Shanxi Provinces of China. The wild resources of *E. mollis* have been sharply reduced and urgent protection is required for the understanding of molecular mechanism.

With the rapid development of the sequencing technologies and decreasing costs associated with them, studies of plastid genomes have increased. Approximately 4354 plant plastid genomes are available in the National Center for Biotechnology Information (NCBI) nucleotide database (NCBI 2020, https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=2759&opt=plastid>), including alimentary crops, economic crops and medicinal plants. In general, plastid genomes of plants are distinguished by having a slower rate of evolution, maternal inheritance, moderate rate of nucleotide replacement and high conservation. Therefore, plastid genomes are widely used in phylogenetic analysis, DNA barcodes and molecular marker development (Jansen *et al.* 2007; Parks *et al.* 2009; Barrett *et al.* 2013). In medical plants, plastid genome sequencing

Yongqin Cheng and Yanci Yang contributed equally to this work.

YC, YY, ZJ and JC conceived and designed the experiment; YC, YY, XF and LL performed the experiments and analysed the data; YC wrote the paper. ZJ and JC revised the article. All the authors read and approved submission of the final manuscript.

provides the possibility for obtaining more abundant DNA molecular information, identifying species of traditional Chinese medicine, and the genetic diversion of genuine medicinal material (Lin *et al.* 2010).

In this study, high-throughput sequencing was used to analyse the plastid genome of *E. mollis*. Additionally, we also performed the comparative genomes of close relatives, which have been published in Elaeagnaceae family (Choi *et al.* 2015; Chen and Zhang 2017), to identify the efficiency of plastid genome for close related species and provide data support for selecting more suitable DNA barcode. Recent studies have shown that the plastid genome sequences are indispensable data for plant phylogenetic (Parks *et al*. 2009). Here, the phylogenetic analysis was constructed using 32 plastid genomes of Rosales, aiming to provide a better understanding of the evolution of *E. mollis*.

Materials and methods

Sample collection and DNA extraction

Samples of *E. mollis* were collected from Xiangning County (N36°02′35.25″, E111°07′57.36″) in Shanxi Province, China. Samples were promptly dried with silica gel using the cetyl trimethyl ammonium bromide (CTAB) method with slight modification. Total genomic DNA was extracted using Plant Genomic DNA Extraction kit (Tiangen, Beijing, China) according to the manufacturer's instructions.

Plastid genomes sequencing, assembly and annotation

The extracted genomic DNA was determined for concentration and quality by QuantiFlour (Promega, USA), and then subjected to high-throughput amplicon sequencing on the Illumina HiSeq 2500 platform (Biomarker Biotechnology, Beijing, China). In addition, we downloaded only two published Elaeagnaceae plastid genomic sequences (*Elaeagnus macrophylla* NC_028066 and *Hippophae rhamnoides* NC_035548) for comparative analysis. Low-quality reads were removed from raw reads using NGSQC Toolkit v. 2.3.3 using default parameters (Patel and Jain 2012). Subsequently, filtered paired-end reads were used to reconstruct the plastid genomes using the program MIRA v. 4.0.2 (Chevreux *et al.* 2004) and MITObim v. 1.7 (Hahn *et al.* 2013). To ensure accurate assembly, *E. macrophylla* (NC_028066) and *H. rhamnoides* (NC_035548) were used as references. The assembled plastid genome sequences were introduced to the program DOGMA (Wyman *et al*. 2004), for annotation and manual correction by comparison with published Elaeagnaceae plastid genome using Geneious R8 (Biomatters, Auckland, New Zealand). Plastid genomes were drawn with OGDRAW (http://ogdraw.mpimp-golm.mpg.de/) (Lohse *et al.* 2013).

Plastid genome comparison and repeat sequence analysis

In this study, a visual alignment of *E. mollis*, *E. macrophylla* and *H. rhamnoides* plastid genomes was generated in mVISTA (Shuffle-LAGAN mode) (Frazer *et al.* 2004). REPuter (Kurtz *et al.* 2001) online software was used to identify the dispersed and palindromic repeat sequences with parameters set as follows: (i) hamming distance of 3, (ii) maximum repeats size of 50 bp, and (iii) and minimum repeat size of 30 bp. Tandem repeat sequences >10-bp long were detected using the online program Tandem Repeats Finder (Benson 1999), with settings of 50 and 500 for the minimum alignment score and maximum period size, respectively. The alignment parameters of match, mismatch, and indel were set as 2, 7 and 7, respectively.

Codon usage bias

In total, 81 protein-coding genes (PCGs) with a length >100 bp were selected for a synonymous codon usage analysis to ensure sampling accuracy. Relative synonymous codon usage (RSCU), the proportion of the observed frequency of a codon to the expected frequency (Sharp and Li 1986), was determined using MEGA 5.0 (Tamura *et al.* 2011). An RSCU value less than 1.0 was considered evidence of a lack of bias, between 1.0 and 1.2 was low bias, between 1.2 and 1.3 was moderated bias, and greater than 1.3 was high bias (Zuo *et al.* 2017).

Phylogenetic analysis

In total, 32 plastid genomes from eight families were downloaded from NCBI GenBank nucleotide database (https://www.ncbi.nlm.nih.gov/). The reliability of phylogenetic analysis is mainly dependent on the accuracy of the sequence alignment (Morrison and Ellis 1997; Ogden and Rosenberg 2006; Hohl and Ragan 2007). Multiple alignments was conducted using the MAFFT program (Katoh and Standley 2013). Then, the best-fitting nucleotide substitution model was selected for the phylogenetic analysis using the ModelGenerator program (Keane *et al.* 2006). The phylogenetic tree was constructed as previously described (Huang *et al.* 2020). Maximum likelihood (ML) analyses were performed using the program RAxML v. 8.1.5 (Stamatakis 2014).

Results

Basic information of the plastid genomes

The plastid genomes obtained in this study for *E. mollis* has been deposited in the GenBank under the accession number

MG386504. It is 151,354 bp long containing one LSC of 81,072 bp and one SSC of 18,560 bp, which were separated by a pair of IRs of 25,861 bp (figure 1). The guanine and cytosine (GC) content of the LSC, SSC and IR regions were 35%, 30.1%, and 42.6%, respectively. In total, 132 genes were predicted in the plastid genomes of *E. mollis*, including eight rRNA genes, 38 tRNA genes, and 86 protein-coding genes. In summary, 113 genes appeared in a single copy and four rRNA genes (*rrn4.5*, *rrn5*, *rrn16* and *rrn23*), eight tRNA genes (*trnA-UGC*, *trnH-GUG*, *trnI-CAU*, *trnI-GAU*, *trnL-CAA*, *trnN-GUU*, *trnR-ACG* and *trnV-GAC*) and seven protein-coding genes (*ndhB*, *rps7*, *rps12*, *rpl2*, *rpl23*, *ycf1* and *ycf2*) appeared in two IRs (table 1). Among the 113 single-copy genes, *ycf1* crossed SSC and IR regions. Additionally, 15 genes (*trnA-UGC*, *trnG-GCC*, *trnI-GAU*, *trnK-UUU*, *trnL-UAA*, *trnV-UAC*, *rpl2*, *rpl16*, *rps12*, *rpoC1*, *atpF*, *petB*, *petD*, *ndhA* and *ndhB*) contained one intron, and two genes (*ycf3* and *clpP*) contained two introns. The longest intron was *trnK-UUU* at 2825 bp (table 2).

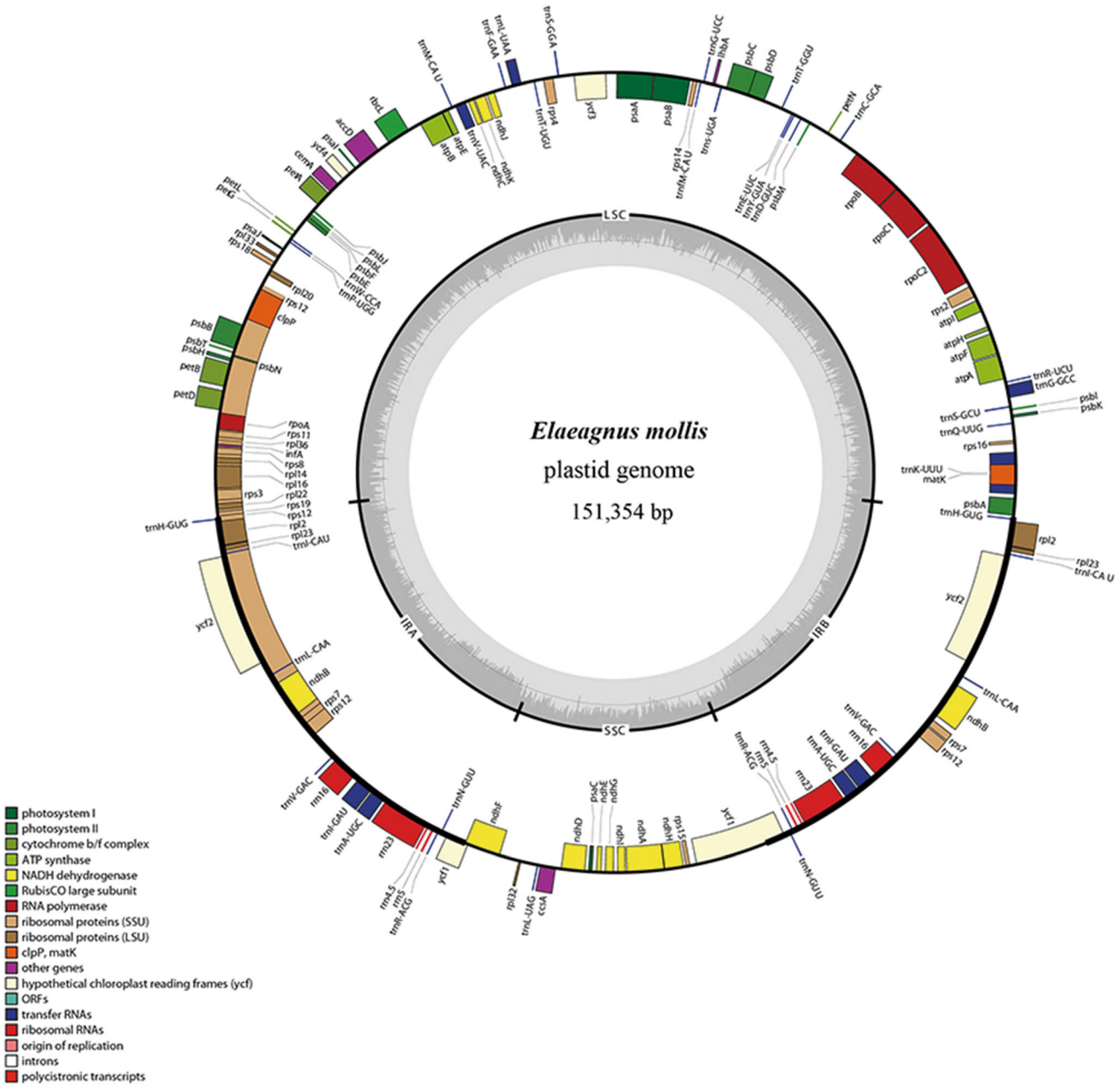

Figure 1. Genetic map of *E. mollis* chloroplast genome. The genes belong to different functional groups are encoded by different colours. The genes outside the circle are transcribed clockwise, while the genes in the circle are transcribed counter clockwise. The inner circle indicates the inverted boundaries and GC content.

Table 1. Genes found in the *E. mollis* plastid genome.

Group of genes	Name of gene	Number
tRNA genes	*trnA-UGC** (×2), *trnC-GCA*, *trnD-GUC*, *trnE-UUC*, *trnF-GAA*, *trnfM-CAU*, *trnG-GCC**, *trnG-UCC*, *trnH-GUG* (×2), *trnI-CAU* (×2), *trnI-GAU** (×2), *trnK-UUU**, *trnL-CAA* (×2), *trnL-UAA**, *trnL-UAG*, *trnM-CAU*, *trnN-GUU* (×2), *trnP-UGG*, *trnQ-UUG*, *trnR-ACG* (×2), *trnR-UCU*, *trnS-GCU*, *trnS-GGA*, *trnS-UGA*, *trnT-GGU*, *trnT-UGU*, *trnV-GAC* (×2), *trnV-UAC**, *trnW-CCA*, *trnY-GUA*	30
Ribosomal RNA genes	*rrn4.5* (×2), *rrn5* (×2), *rrn16* (×2), *rrn23* (×2)	4
Large subunit of ribosomal proteins	*rpl2** (×2), *rpl14*, *rpl16**, *rpl20*, *rpl22*, *rpl23* (×2), *rpl32*, *rpl33*, *rpl36*	9
Small subunit of ribosomal proteins	*rps2*, *rps3*, *rps4*, *rps7* (×2), *rps8*, *rps11*, *rps12** (×2), *rps14*, *rps15*, *rps16*, *rps18*, *rps19*	12
DNA-dependent RNA polymerase	*rpoA*, *rpoB*, *rpoC1 **, *rpoC2*	4
Translation initiation factor	*infA*	1
Large subunit of rubisco	*rbcL*	1
Photosystem I	*psaA*, *psaB*, *psaC*, *psaI*, *psaJ*	5
Photosystem II	*psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbN*, *psbT*	14
ATP synthase	*atpA*, *atpB*, *atpE*, *atpF**, *atpH*, *atpI*	6
Cytochrome b/f complex	*petA*, *petB**, *petD**, *petG*, *petL*, *petN*	6
NADPH dehydrogenase	*ndhA**, *ndhB** (×2), *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *ndhJ*, *ndhK*	11
Maturase	*MatK*	1
Protease	*clpP***	1
Envelope membrane protein	*cemA*	1
Subunit acetyl-CoA-carboxylase	*accD*	1
c-Type cytochrome synthesis gene	*ccsA*	1
Conserved open reading frames (*ycf*)	*ycf1* (×2), *ycf2* (×2), *ycf3*, *ycf4*	4
Genes of unknown function	*lhbA*	1
Total	113	

*Gene with one intron; **genes with two introns; (×2) genes with two copies.

Comparison with plastids in other Elaeagnaceae species

We compared the basic features of three plastid genomes obtained in our present study with two previously published Elaeagnaceae plastid genomes (*E. macrophylla* and *H. rhamnoides*). Similarly, each of the three plastid genomes contained 38 tRNA genes and eight rRNA genes. The length of the three genomes ranged from 151,354 bp (*E. mollis*) to 156,123 bp (*H. rhamnoides*). The lengths of the LSC region varied from 81,072 bp (*E. mollis*) to 83,331 bp (*H.*

Table 2. Genes with introns in the *E. mollis* plastid genome and length of exons and introns.

Gene name	Location	Exon I (bp)	Intron I (bp)	Exon II (bp)	Intron II (bp)	Exon III (bp)
trnA-UGC	IR	38	803	35		
trnG-GCC	LSC	23	689	48		
trnI-GAU	IR	42	951	35		
trnK-UUU	LSC	35	2482	37		
trnL-UAA	LSC	37	507	50		
trnV-UAC	LSC	37	589	39		
rpl2	IR	435	662	390		
rpl16	LSC	402	988	9		
rps12	IR	26	535	232		
rpoC1	LSC	1593	706	435		
atpF	LSC	411	724	144		
petB	LSC	6	789	642		
petD	LSC	8	688	475		
ndhA	SSC	540	1185	552		
ndhB	IR	756	680	777		
clpP	LSC	228	595	289	827	6879
ycf3	LSC	153	694	228	751	4293

rhamnoides) (table 3). SSC region lengths ranged from 18,560 bp (*E. mollis*) to 18,831 bp (*H. rhamnoides*), and IR lengths varied from 25,861 bp (*E. mollis*) to 26,658 bp (*H. rhamnoides*) (table 1). The GC contents of plastid genomes of *E. mollis*, *E. macrophylla* and *H. rhamnoides* were 37.0%, 41.1% and 36.7%, respectively (table 3).

Forty-seven repeat sequences were identified in the plastid genomes of *E. mollis*, including tandem (17), dispersed (12), and palindromic (18) types. The distribution of the repetitive sequences in the plastid genomes of *E. mollis*, *E. macrophylla* and *H. rhamnoides* were similar but the dispersed repeats was the least (figure 2).

Contraction and expansion at the boundaries of IR region is a common evolutionary event that largely accounts for the variation in angiosperm plastid genomes sizes. These events play a crucial role in evolution (Kode *et al.* 2005; Raubeson *et al.* 2007; Yao *et al.* 2015). Genomic structure and size of the three plastids were highly conserved, the IR/SSC boundary regions varied slightly (figure 3). The border between the IRB and SSC encompassed the *ycf1* gene, with *ycf1* pseudogenes found in plastids of all three species. The length of the *ycf1* pseudogene was similar in the three Elaeagnaceae species. Overlap between *ndhF* and *ycf1* was noted in all three Elaeagnaceae species (*E. mollis*, *E. macrophylla* and *H. rhamnoides*), with *ndhF* expanding into the IRB region by 10 bp, 11 bp and 7 bp, respectively.

The SSC–IRA junctions were situated in the *ycf1* coding region and the length of *ycf1* in the IRA region varied in three species from 1215 bp to 1247 bp. The IRA–LSC junctions were located between *trnH* and *psbA*, and at the IRA–LSC junction, the *trnH* gene extended 83 bp, 118 bp, and 214 bp into the IRA region in *E. mollis*, *E. macrophylla* and *H. rhamnoides*, respectively. In summary, genes other than *rps19* in these three species showed similar compositions at the IR-SSC and IR-LSC boundaries (figure 3).

The *psbA* gene was located in the LSC region in all three species. The *rps19* genes in *E. macrophylla* and *E. mollis* were all found in the LSC region, whereas the IR region extended into a short *rps19* pseudogene of 74 bp in *H. rhamnoides*. The *trnH-GUG* genes were all located in the IR region, and their distances from the LSC-IRB boundary were 85 bp, 118 bp, and 74 bp and from the LSC-IRA boundary were 83 bp, 118 bp, and 214 bp in *E. mollis*, *E. macrophylla*, and *H. rhamnoides*.

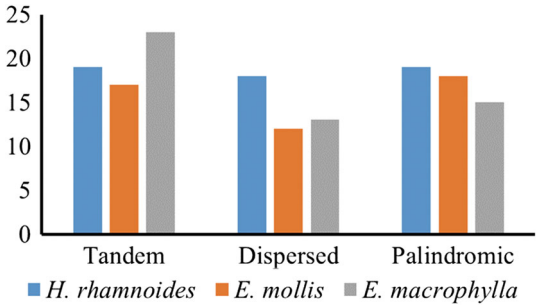

Figure 2. Numbers of repeats in the plastid genome of *E. mollis* compared with other two species.

An analysis of plastid genomes is crucial for understanding the relatedness of plant species and increasingly valuable with more published DNA sequences. There have been numerous comparative analyses of plastid genomes to examine interspecific relationships and identify specific DNA barcodes in closely related plant species (Chen *et al.* 2012). The sequences of plastid genomes from *E. mollis* were compared with those from *E. macrophylla* and *H. rhamnoides* using the mVISTA program. The results showed a few substantial differences between the plastid genomes of the three Elaeagnaceae species (figure 4). These differences, which could be used as specific DNA barcodes, emerged in the intergenic regions of *atpH-atpI*, *petN-psbM*, *trnT-psbD*, *trnP-psaJ*, *rpl32-trnL* and *ycf1*. The tRNA and rRNA coding regions (light blue) were the most highly conserved, and figure 4 shows that conserved noncoding sequences (CNS) were more divergent than coding region (exon). We also found *trnH-GUG* duplication in the IR regions of plastids in the three species.

Codon usage bias

The codon usage frequency in plastids genomes of *E. mollis* is shown in table 4. A high degree of conservation is apparent. Leucine (Leu, ~10.52%) and cysteine (Cys, ~1.4%) were the most and least used amino acids,

Table 3. The features of plastid genomes of three Elaeagnaceae species.

Species	*E. mollis*	*E. macrophylla*	*H. rhamnoides*
Genome length	151,354	152,224	156,123
LSC length	81,072	82,136	83,970
SSC length	18,560	18,278	18,831
IR	25,861	25,905	26,658
Total gene number	132	131	132
No. of protein coding genes	86	85	86
No. of tRNA genes	38	38	38
No. of rRNA genes	8	8	8
GC content in genome (%)	37.0	41.1	36.7

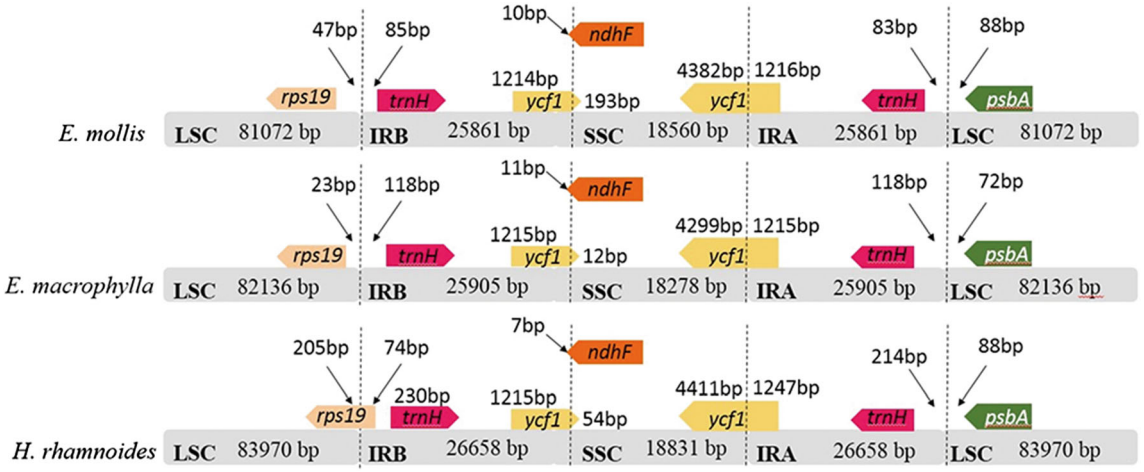

Figure 3. Comparison of four junctions (LSC/IRB, IRb/SSC, SSC/IRA and IRA/LSC) among three Elaeagnaceae genome.

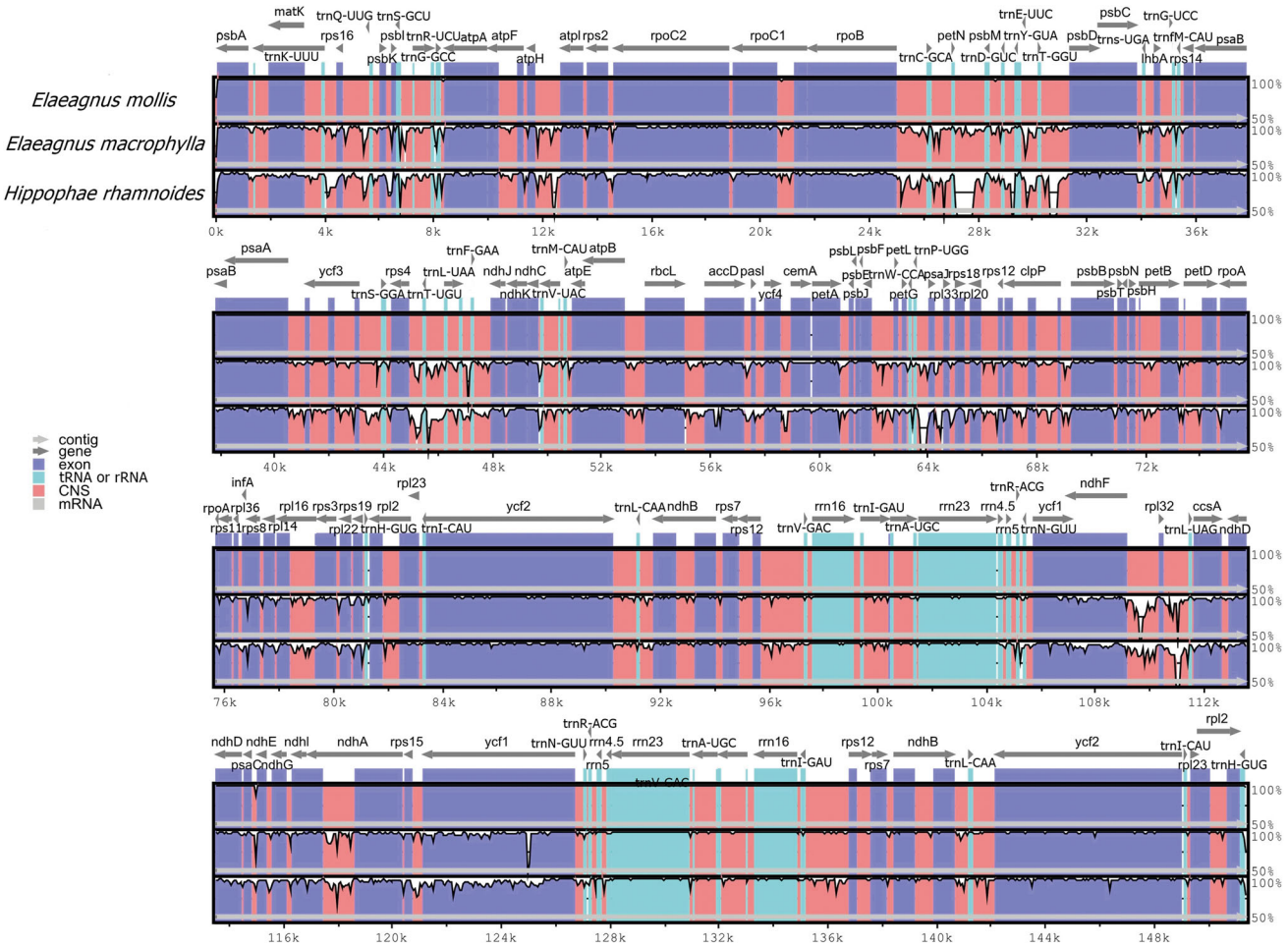

Figure 4. mVISTA per cent identity plot comparing the three chloroplast genomes with *E. mollis* as a reference. The *y*-axis represents the per cent identity within 50–100%. Grey arrows and thick black lines above the alignment indicate genes with their orientation and the position of IRs. The purple regions represent exons, the light-blue regions represent rRNA or tRNA coding genes, the pink regions represent CNS, the gray regions represent mRNA, and white peaks represent differences of genomics.

Table 4. Synonymous codon usage in the *E. mollis* plastid genome.

Amino acid	Abbreviation	Codon	Count	RSCU	AA frequency (%)	Amino acid	Abbreviation	Codon	Count	RSCU	AA frequency (%)
Ala	A	GCG	149	0.44	5.18	Pro	P	CCA	302	1.12	4.08
		GCA	392	1.15				CCC	229	0.85	
		GCC	207	0.61				CCG	135	0.5	
		GCU	617	1.81				CCU	410	1.52	
Cys	C	UGC	71	0.47	1.14	Gln	Q	CAA	719	1.53	3.56
		UGU	230	1.53				CAG	218	0.47	
Asp	D	GAC	197	0.38	3.91	Arg	R	CGA	345	1.37	5.73
		GAU	833	1.62				CGC	98	0.39	
Glu	E	GAA	1059	1.52	5.30			CGG	105	0.42	
		GAG	338	0.48				CGU	322	1.28	
Phe	F	UUC	523	0.67	5.89			AGA	472	1.87	
		UUU	1028	1.33				AGG	169	0.67	
Gly	G	GGG	296	0.67	6.76	Ser	S	UCA	399	1.19	7.61
		GGA	725	1.63				UCC	324	0.97	
		GGC	165	0.37				UCG	184	0.55	
		GGU	594	1.33				UCU	573	1.71	
His	H	CAC	152	0.47	2.43			AGC	122	0.37	
		CAU	488	1.52				AGU	403	1.21	
Ile	I	AUA	737	0.97	8.68	Ter		UAA	49	1.79	0.35
		AUC	453	0.59				UAG	20	0.73	
		AUU	1098	1.44				UGA	13	0.48	
Lys	K	AAG	357	0.49	5.55	Thr	T	ACA	427	1.28	5.07
		AAA	1106	1.51				ACC	243	0.73	
Leu	L	UUG	569	1.23	10.52			ACG	146	0.44	
		UUA	865	1.87				ACU	521	1.56	
		CUG	189	0.41		Val	V	GUA	535	1.52	5.35
		CUA	375	0.81				GUC	162	0.46	
		CUC	181	0.39				GUG	193	0.55	
		CUU	593	1.28				GUU	519	1.47	
Met	M	AUG	614	1	2.33	Trp	W	UGG	474	1	1.80
Asn	N	AAC	318	0.48	5.02	Tyr	Y	UAC	203	0.41	3.78
		AAU	1005	1.52				UAU	793	1.59	

respectively. Among the 20 amino acids, methionine (Met) and tryptophan (Trp) were encoded by only one codon, whereas the others were encoded by multiple codons. Isoleucine (Ile) were encoded by three codons, AUA, AUC and AUU, and their corresponding RSCU values were 0.97, 0.59, and 1.44, respectively. Therefore, AUU was the preferred translation codon. Bias in the use of synonymous codons for amino acids other than Met and Trp, which were similar in plastid genomes of higher plants. The results indicated that the third nucleotide of the preferred translation codons were mostly A or U.

Phylogenetic analysis of Rosales

Plastid genome sequences have been extensively used to analyse the plant phylogenies (Goremykin *et al.* 2015; Sun et al. 2016). Phylogenetic trees were reconstructed using ML methods in 32 species from the order Rosales. *Castanea mollissima* was set as outgroup to infer the phylogenetic positions of *Elaeagnus* within Elaeagnaceae and relationships of Rosales. Whole plastid genomes were used to construct the phylogenetic trees.

Except for the small family Barbeyaceae and no published sequence of Dirachmaceae, most Rosales families were represented by three to five samples. As shown in figure 5, Rosales families were divided into two clades with strongly supported. Rosaceae in the first clade were sister to the rest of the order. The remaining Rosales were divided into two subclades: (i) Ulmaceae, Cannabaceae, Boehmeria and Moraceae (MLBS=100%); (ii) Barbeyaceae, Elaeagnaceae, and Rhamnaceae (MLBS=100%). In the first of these subclades, Boehmeria and Moraceae were well-supported as monophyletic (MLBS=100%); Boehmeria +Moraceae, in turn, formed a clade with Cannabaceae (MLBS=100%); finally, Ulmaceae were sister to Cannabaceae, Moraceae and Urticaceae (MLBS=100%). In the second of these subclades, Rhamnaceae were sister to the rest of this clade, which formed a well-supported monophyletic group (MLBS =100%). Barbeyaceae plus Elaeagnaceae formed a clade with Rhamnaceae (MLBS=87%).

During the evolution of plants, many genes have been extirpated from plastid genomes. By far, *infA* is the most mobile plastid gene in plants. A study by Millen *et al.* (2001) suggests that many *infA* copies in plastid DNA was lost during angiosperm evolution. Our results showed that *infA*

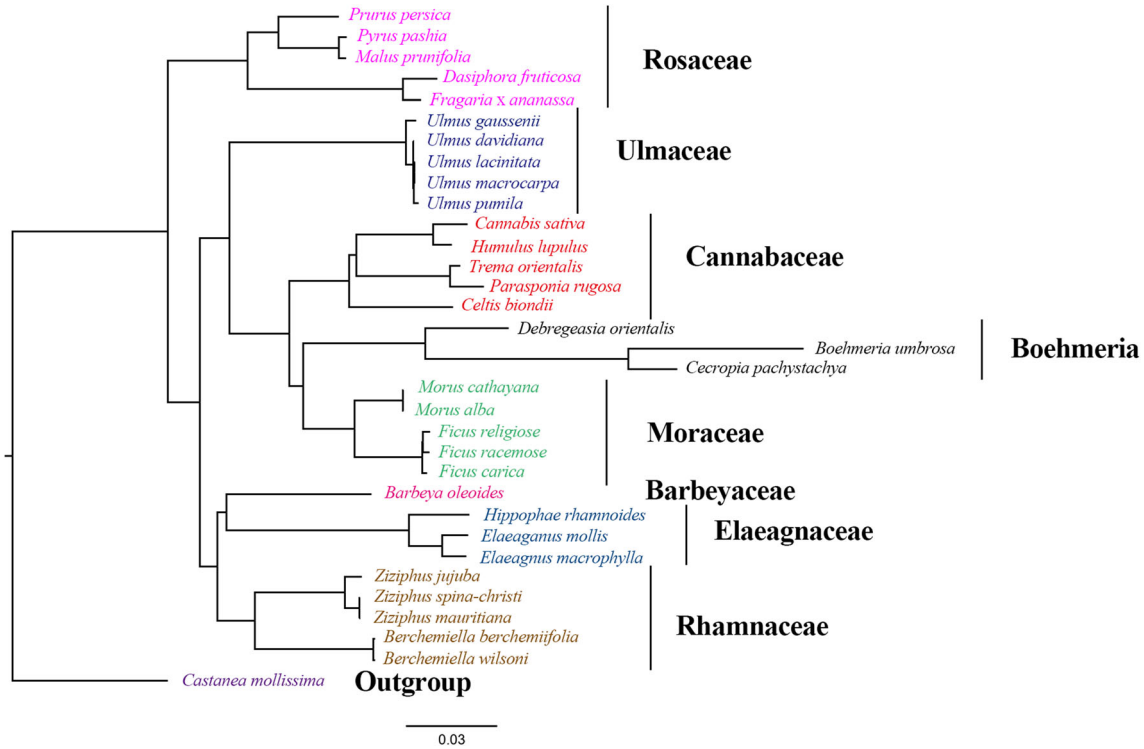

Figure 5. Phylogenetic relationships based on complete chloroplast genome sequences of species belonging to Rosales inferred from ML analysis.

gene has been lost from Rosaceae, Ulmaceae, Moraceae and four genera of Rhamnaceae. Moraceae, Ulmaceae, Barbeyaceae, Cannabaceae, Boehmeria, Rosaceae and partial species of Rhamnaceae. Further, comparative analysis of Rosales plastid genomes indicated that the duplication of *trnH* only occurs in Elaeagnaceae.

Discussion

Plastid genomes of *E. mollis* were similar to the previous studies of the plastid genome of *E. mollis* (NC_036932) in size, structure, gene content, GC content and typical quadripartite structure (Wang *et al*. 2017a). In this study, plastid genomes of *E. mollis* were compared with *E. macrophylla* and *H. rhamnoides*, we found that all the plastid genomes possessed the typical quadripartite structure with circular and double-stranded DNA. The size, structure, gene content, and GC content of the newly generated plastid genomes were generally similar to those published for Elaeagnus, which revealed that plastid genomes are conserved in Elaeagnus. Additionally, six divergence hotspots (*atpH-atpI*, *petN-psbM*, *trnT-psbD*, *trnP-psaJ*, *rpl32-trnL* and *ycf1*) were reported by comparing the plastid genomes of the three Elaeagnaceae species, which could be used as molecular genetic markers for population genetics and phylogenetic studies. Although *infA* gene is always absent in most Rosales plants (Millen *et al*. 2001; Su *et al*. 2014; Choi *et al*. 2015), it was intact in Elaeagnus. Compared with other plants of the same genus, the IR region of plastid genome also changed slightly, which may be caused by the contraction and expansion of IR region (Huang *et al*. 2014). Codon usage bias reflects synonymous codons that have different usage frequencies (Ermolaeva 2001). The results showed a strong bias toward A/T at the third codon position, which is in line with previous findings of a third-nucleotide preference for A/T in other land plants (Wang *et al*. 2017b; Zhou *et al*. 2017).

The duplication of *trnH* gene only occurs in Elaeagnaceae, which could be a useful marker in Rosales. Elaeagnaceae include three genera, i.e. *Elaeagnus*, *Hippophae* and *Shepherdia*. Elaeagnaceae have been included in different orders in different classifications, such as close to Proteaceae, Rhamnaceae, Thymelaeaceae or Penaeaceae (Jansen *et al*. 2000). The phylogenetic analysis of Rosales revealed that Elaeagnaceae are sister to Barbeyaceae. Elaeagnaceae in a clade composed of Barbeyaceae and Rhamnaceae, Rosaceae were sister to other Rosales, which were similar with previous studies (Sytsma *et al*. 2002; Wang *et al*. 2009; Zhang *et al*. 2011). The remainder of the order comprises two subclades; (i) Ulmaceae are sister to Cannabaceae plus (Boehmeria and Moraceae); (ii) Rhamnaceae are sister to Elaeagnaceae plus Barbeyaceae. The plastid genome sequences fully resolved phylogenetic relationships within Rosales with strong internal support. This is the first analysis of complete plastids of all Rosales lineages,

and our results are generally similar to the previous observations which used two nuclear and 10 plastid loci to infer phylogeny of Rosales (Zhang *et al.* 2011).

Conclusions

In this study, we successfully sequenced plastid genomes of *E. mollis* and compared them with plastids from other species of Elaeagnaceae to determine the sequence variations and molecular phylogenies. Plastids in *E. mollis* showed a typical quadripartite DNA molecular structure, which is similar to those in other angiosperm species. We determined divergence hotspots of medicinal and economic values in *E. mollis* that could be used as potential genetic markers for the further studies.

We reconstructed the phylogenetic relationships of Rosales based on the complete plastid genomic data, aiming to clarify the evolutionary relationships among the major clades of Rosales. We successfully resolved the evolutionary relationships among the major clades of Rosales and also found that the duplication of *trnH* occurs only in Elaeagnaceae, suggesting that *trnH* may be an important marker for the phylogenetic studies of Rosales. Our study not only contributes to a basic understanding of *E. mollis* plastid genomes and provides a valuable resource for the evolutionary research in the Elaeagnaceae family, but also demonstrates the effectiveness of plastid phylogenomics to further resolve the phylogenetic relationships in Rosales.

Acknowledgements

The authors express gratitude to all those who helped while writing this thesis. The study was funded by the Special Fund for Forest Scientific Research in the Public Welfare (201204308).

References

Barrett C. F., Davis J. I., Leebens-Mack J., Conran J. G. and Stevenson D. W. 2013 Plastid genomes and deep relationships among the commelinid monocot angiosperms. *Cladistics* **29**, 65–87.

Benson G. 1999 Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. **27**, 573–580.

Chen S. L., Xu J., Liu C., Zhu Y. J., Nelson D. R., Zhou S. G. *et al*. 2012 Genome sequence of the model medicinal mushroom *Ganoderma lucidum*. *Nat. Commun.* **3**, 913.

Chen S. Y. and Zhang X. Z. 2017 Characterization of the complete chloroplast genome of seabuckthorn (*Hippophae rhamnoides* L.). *Conserv. Genet. Resour*. **9**, 1–4.

Chevreux B., Pfisterer T., Drescher B., Driesel A. J., Müller W. E., Wetter T. *et al*. 2004 Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res*. **14**, 1147–1159.

Choi K. S., Son O. G. and Park S. J. 2015 The chloroplast genome of *Elaeagnus macrophylla* and *trnH* duplication event in Elaeagnaceae. *PLoS One* **10**, e0138727.

Ermolaeva M. D. 2001 Synonymous codon usage in bacteria. *Curr. Issues Mol. Biol.* **3**, 91–97.

Frazer K. A., Pachter L., Poliakov A., Rubin E. M. and Dubchak I. 2004 VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* **32**, W273–W279.

Goremykin V. V., Nikiforova S. V., Cavalieri D., Pindo M. and Lockhart P. 2015 The root of flowering plants and total evidence. *Syst. Biol.* **64**, 879–891.

Huang H., Shi C., Lui Y., Mao S.Y. and Gao L. Z. 2014 Thirteen Camellia chloroplast genome sequences determined by high-throughput sequencing: genome structure and phylogenetic relationships. *BMC Evol. Biol*. **14**, 151.

Hahn C., Bachmann L. and Chevreux B. 2013 Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads-a baiting and iterative mapping approach. *Nucleic Acids Res.* **41**, e129.

Hohl M. and Ragan M. A. 2007 Is multiple-sequence alignment required for accurate inference of phylogeny? *Syst. Biol.* **56**, 206–221.

Huang Z., Wang D.D, Li J.P., Wei C. and He H. 2020 Transovarial transmission of bacteriomes-associated symbionts in the cicada Pycna repanda (Hemiptera: Cicadidae). *Appl. Environ. Microbiol.* **86**, e02457-19.

Jansen R. K., Cai Z. Q., Raubeson L. A., Daniell H., Depamphilis C. W., Leebens-Mack J. *et al*. 2007 Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proc. Natl. Acad. Sci. USA* **104**, 19369–19374.

Jansen S., Piesschaert F. and Smets E. 2000 Wood anatomy of Elaeagnaceae, with comments on vestured pits, helical thickenings, and systematic relationships. *Am. J. Bot.* **87**, 20–28.

Katoh K. and Standley D. M. 2013 MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780.

Keane T. M., Creevey C. J., Pentony M. M., Naughton T. J. and Mclnerney J. O. 2006 Assessment of methods for amino acid matrix selection and their use on empirical data shows that *ad hoc* assumptions for choice of matrix are not justified. *BMC Evol. Biol*. **6**, 29.

Kode V., Mudd E. A., Iamtham S. and Day A. 2005 The tobacco plastid *accD* gene is essential and is required for leaf development. *Plant J*. **44**, 237–244.

Kurtz S., Choudhuri J. V., Ohlebusch E., Schleiermacher C., Stoye J. and Giegerich R. 2001 REPuter: The manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res*. **29**, 4633–4642.

Liang S. H., Yang R. N., Dong C. W. and Yang Q. P. 2015 Physicochemical properties and fatty acid profiles of *Elaeagnus mollis* Diels nut oils. *J. Oleo. Sci.* **64**, 1267–1272.

Lin X. H., Liu Z. H., Li Q., Chen S. L. and Li X. W. 2010 Chloroplast genome of medicinal plants. *World Sci. Tech.* (*Modernization of Traditional Chinese Medicine and Materia Medica*) **3**, 139-143.

Lohse M., Drechsel O., Kahlau S. and Bock R. 2013 Organellar Genome DRAW-a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Res.* **41**, 575–581.

Millen R. S., Olmstead R. G., Adams K. L., Palmer J. D., Lao N. T. and Heggie L. *et al*. 2001 Many parallel losses of *infA* from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. *Plant Cell*. **13**, 645–658.

Morrison D. A. and Ellis J. T. 1997 Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of apicomplexa. *Mol. Biol. Evol.* **14**, 428–441.

Ogden T. H. and Rosenberg M. S. 2006 Multiple sequence alignment accuracy and phylogenetic inference. *Syst. Biol.* **55**, 314–328.

Parks M., Cronn R. and Liston A. 2009 Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biol.* **7**, 84.

Patel R. K. and Jain M. 2012 NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One* **7**, e30619.

Raubeson L. A., Peery R., Chumley T. W., Dziubek C., Fourcade H. M., Boore J. L. *et al*. 2007 Comparative chloroplast genomics: Analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. *BMC Genom*. **8**, 174.

Sharp P. M. and Li W. H. 1986 An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol*. **24**, 28–38.

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

Su H. J., Hogenhout S. A., Al-Sadi A. M. and Kuo C. H. 2014 Complete Chloroplast Genome Sequence of Omani Lime (*Citrus aurantiifolia*) and Comparative Analysis within the Rosids. *PLoS One* **9**, e113049.

Sun L. H., Fang L., Zhang Z. H., Chang X., Penny D. and Zhong B. J. 2016 Chloroplast phylogenomic inference of green algae relationships. *Sci. Rep*. **6**, 20528.

Sytsma K. J., Morawetz J., Pires J. C., Nepokroeff M., Conti E., Zjhra M. *et al*. 2002 Urticalean rosids: circumscription, rosid ancestry, and phylogenetics based on rbcL, trnL-F, and ndhF sequences. *Am. J. Bot*. **89**, 1531–1546.

Tamura K. J., Morawetz J., Pires J. C., Nepokroeff M., Conti E., Zjhra M. *et al*. 2011 MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol*. **28**, 2731–2739.

Wang W. C., Chen S. Y. and Zhang X. Z. 2017a Characterization of the complete chloroplast genome of *Elaeagnus mollis*, a rare and endangered oil plant. *Conserv. Genet. Resour.* **9**, 439–442.

Wang W. B., Yu H., Wang J. H., Lei W. J., Gao J. H., Qiu X. P. *et al.* 2017b The complete chloroplast genome sequences of the medicinal plant *Forsythia suspensa* (Oleaceae). *Int. J. Mol. Sci.* **18**, 2288.

Wang H. C., Moore M. J., Soltis P. S., Bell C. D., Brockington S. F., Alexandre R. *et al*. 2009 Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proc. Natl. Acad. Sci. USA* **106**, 3853–3858.

Wyman S. K., Jansen R. K. and Boore J. L. 2004 Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* **20**, 3252–3255.

Xie S. L. and Ling Y. J. 1997 The biology features and conservation of *Elaeagnus mollis*, a rare and endangered plant species. *Bull. Bot. Res.* **17**, 153–157.

Yao J. X. 2005 Effective protection and rational development in precious tree species, *Elaeagnus mollis. Shanxi Forest. Sci. Tech*. **1**, 39-41.

Yao X. H., Tang P., Li Z. Z., Li D. W., Liu Y. F. and Huang H.W. 2015 The first complete chloroplast genome sequences in Actinidiaceae: genome structure and comparative analysis. *PLoS One* **10**, e0129347.

Zhang J. M. and Zhang F. 2015 Population structure and genetic variation of the endangered species *Elaeagnus mollis* Diels (Elaeagnaceae). *Genet. Mol. Res*. **14**, 5950–5957.

Zhang S. D., Soltis D. E., Yang Y., Li D. Z. and Yi T. S. 2011 Multi-gene analysis provides a well-supported phylogeny of Rosales. *Mol. Phylogenet. Evol.* **60**, 21-28.

Zhou J. G., Chen X. L., Cui Y. X., Sun W., Li Y. H., Wang Y. *et al*. 2017 Molecular structure and phylogenetic analyses of complete chloroplast genomes of two *Aristolochia* medicinal species. *Int. J. Mol. Sci*. **18**, 1839.

Zuo L. H., Shang A. Q., Zhang S., Yu X. Y., Ren Y. C. *et al*. 2017 The first complete chloroplast genome sequences of *Ulmus* species by de novo sequencing: Genome comparative and taxonomic position analysis. *PLoS One* **12**, e0171264.

Corresponding editor: H. A. Ranganath