Botanica SERBICA

(2023) 47 (2): 279–290

DOI: https://doi.org/10.2298/BOTSERB2302279X
journal homepage: botanicaserbica.bio.bg.ac.rs

Original Scientific Paper

The complete chloroplast genome sequence of *Rhododendron fortunei*: structural comparative and phylogenetic analysis in the Ericaceae family

Yunli Xiao[1], Wanjing Zhang[1], Yirong Sun[1], Zhiliang Li[1], Jiaojun Yu[1], Chunyu Zhang[1,2] and Shuzhen Wang[1,*]

1 College of Biology and Agricultural Resources, Huanggang Normal University, Huanggang, 438000, Hubei Province, China
2 College of Plant Science & Technology, Huazhong Agricultural University, Wuhan, 430070, Hubei Province, China
* Correspondence: wangshuzhen710@whu.edu.cn

ABSTRACT:
Rhododendron fortunei (Ericaceae) possesses valuable horticultural and medicinal values. However, the genomic information on *R. fortunei* is very limited. In this study, the complete chloroplast genome (cp) of *R. fortunei* was assembled and annotated, SSR loci were characterised, comparative genomic analysis was carried out, and phylogenetic research was also performed. The results showed that the *R. fortunei* cp genome was of a typical quadripartite structure (200,997 bp). The lengths of the large single copy region (LSC), the inverted repeat regions (IR), and the small single copy region (SSC) were 109,151 bp, 2,604 bp, and 44,619 bp, respectively. A total of 147 unique genes were identified, including 99 protein-coding genes, 42 tRNA genes, and 6 rRNA genes, respectively. Leucine (11.51%) and cysteine (1.15%) were the highest and lowest representative amino acids, respectively. The total of 30 codons with obvious codon usage bias were all A/U-ending codons. Among the 77 simple sequence repeats, the majority were mononucleotide A/T repeats located in the intergenic spacer region. Five gene regions showed high levels of nucleotide diversity (Pi > 0.03). The comparative genome analysis revealed 7 hotspot intergenic regions (*trnI-rpoB*, *trnT-rpl16*, *rpoA-psbJ*, *rps7-rrn16*, *ndhI-rps16*, *rps16-rps19*, and *rrn16-trnI*), showing great potential as molecular makers for species authentication. Expansion and contraction were detected in the IR region of the *R. fortunei* cp genome. In the phylogenetic tree, *R. fortunei* was closely related to *R. platypodum*. This research will be beneficial for evolutionary and genetic diversity studies of *R. fortunei* and related species among the Ericaceae family.

Keywords:
Rhododendron fortunei, next-generation sequencing, chloroplast genome, comparative genomics, conservation genetics

UDC: 575.86:582.688.3

Received: 31 January 2023
Revision accepted: 04 June 2023

INTRODUCTION

The *Rhododendron* genus (Ericaceae), consisting of more than 1000 species and a large number of vascular plants, is widely distributed around the northern hemisphere (Popescu & Kopp 2013). *Rhododendron fortunei* Lindl, a member of the *Rhododendron* genus mainly distributed in central and south China, is famous for beautiful vegetative forms and bright-coloured flowers (Wang *et al*. 2018). As an evergreen shrub or small tree, *R. fortunei* possesses pink flowers. Furthermore, *R. fortunei* is an important member in forest ecology, and plays critical roles in ecological balance (Wang *et al*. 2018). Nowadays, wild *Rhododendron* populations are greatly affected by habitat fragmentation and *Rhododendron*-based ecological tourism (Wang *et al*. 2019). Therefore, research on the genetic diversity, population evolution, and ecological conservation of wild *Rhododendron* is essential. However, very limited genome data is available regarding *R. fortunei*, which has greatly impeded corresponding research.

Plant plastomes exhibit a circular and quadripartite architecture, including two inverted repeat regions (IRa and IRb), a large single-copy region (LSC), and a small single-copy region (SSC) (Abdullah Henriquez *et al.* 2021). In total, 130 genes related to photosynthesis and carbon fixation have been identified in plant cp genomes (Daniell *et al.* 2016). Due to their conserved gene content and stable structure, the maternally inherited plastomes (107–280 kb) serve as important genetic markers for genome-wide evolutionary investigation (Smith 2015; Zhang *et al.* 2017; Givnish *et al.* 2018; Rossini *et al.* 2021). The complete chloroplast genome (cp genome), rich in evolutionary information, has provided significant genetic information and molecular markers useful for resolving obscure phylogenetic relationships in higher plants (Luo *et al.* 2014; Yap *et al.* 2015). Although the substitution rate of the cp genome is relatively lower than that of the nuclear genome, certain genes exhibit accelerated evolution rates, including *matK*, *ycf1*, and *rbcL* (Dong *et al.* 2015; Wambugu *et al.* 2015).

Next-generation sequencing (NGS), i.e. revolutionised genomic and transcriptomic approaches to biology, has significantly increased the availability of genome data for model and non-model species, as well as facilitating comparative genomics and phylogenetic studies at the interspecific level (Santos & Almeida 2019; Yu *et al.* 2022). In this study, the complete cp genome of *R. fortunei* was assembled and annotated, comparative genomics and phylogenetic research was carried out, and SSR loci were also characterised in the aim of providing a genetic resource for *R. fortunei* and related species in the Ericaceae family.

MATERIALS AND METHODS

Plant materials and DNA extraction. Fresh young leaves of *R. fortunei* were collected from the Dabie Mountains (central China, N 28.76°, E 115.84°, 998 m a.s.l.). The sample collection was authorised by the Department of Biodiversity Conservation at Huanggang Normal University. The specimen identification was carried out by Dr. Hongjin Dong (PhD in Botany). After being immediately dried in silica, the plant materials were stored at -20°C for further usage. The total genomic DNA was extracted and purified according to Wang *et al.* (2019). Furthermore, the quality of the total DNA was verified in 1% agarose gel and quantified using a spectrophotometer (NanoDrop 1000, Thermofisher Scientific, USA). All the plant materials were well conserved in Huanggang Normal University Herbarium (Hubei province, China).

Library construction, genome sequencing, assembly, and annotation. Paired-end Illumina libraries were constructed with NEBNext DNA Library Prep Kit following the manufacturer's recommendations. The randomly fragmented genomic DNA fragments (350 bp) were end polished, A-tailed, and ligated with the NEBNext adapter for Illumina sequencing and PCR enriched. After purification, these obtained libraries were sequenced using the Illumina NovaSeq6000 Sequencing System (Hayward, CA) in a paired-end run (500 cycles, 2 × 150 pb). PRINSEQlite v0.20.4 was used to obtain raw reads, which were subjected to *de novo* assembly with SPAdes software (v 3.15.4) (Schmieder & Edwards 2011; Hanussek *et al.* 2021). Furthermore, PCR amplification and Sanger sequencing were used to close any gaps (Dong *et al.* 2013). The physical map of the *R. fortunei* cp genome was drawn using the online programme Organelle Genome DRAW (OGDRAW v1.3.1) (Greiner *et al.* 2019). Gene annotation and analysis were carried out with CPGAVAS2 software (Liu *et al.* 2012). Finally, the cp genome of *R. fortunei* was submitted to the National Centre for Biotechnology Information (NCBI) (https://www.ncbi. nlm.nih.gov/).

Codon usage and simple sequence repeats analysis. All the protein coding genes of the *R. fortunei* cp genome were kept for analysis. CodonW (v1.4) software (https://sourceforge.net/projects/codonw/) was used for codon usage frequency analysis. Furthermore, relative synonymous codon usage (RSCU) analysis was performed to validate codon usage bias, as the RSCU value was defined as the ratio of the observed frequency of codons to the expected frequency regarding the equal usage of synonymous codons for a certain amino acid (Rossini *et al.* 2021; Shi *et al.* 2023). The codons with RSCU values over 1 were considered as preferred codons (Morton 2022).

The *R. fortunei* cp genome sequences were screened using MISA online software (v2.1) (MicroSAtellite, http://pgrc.ipk-gatersleben.de/misa) for searching SSR motifs (Thiel *et al.* 2003). The minimum numbers of repeat units were set as follows: 10 repeat units for mono-nucleotide SSRs; five repeat units for di-nucleotide SSRs; four repeat units for tri-nucleotide SSRs; three repeat units for penta-, tetra-, and hexanucleotide SSRs. For compound SSRs, the minimum distance between two SSR loci was set as 100 bp.

Nucleotide diversity analysis and genome comparative analysis. In total, 9 available cp genomes of the *Rhododendron* genus were downloaded from the NCBI database, including *R. pulchrum* (MN182619), *R. delavayi* (MN413198), *R. henanense* (MT239363.1), *R. micranthum* (MT239365), *R. concinnum* (MT239366), *R. riersonianum* (MT533181), *R. simsii* (MW030509), *R. molle* (MZ073672), and *R. platypodum* (NC_053746). All these cp genomes were pooled, and unique genes were extracted and aligned by PhyloSuite v1.2.2 and MAFFT v7 software, respectively. DnaSP 6.12.03 software was used to calculate the nucleotide diversity (Pi) (Rossini *et al.* 2021).

Fig. 1. The gene maps of the complete cp genome of *Rhododendron fortunei*. The thick lines represent the large single-copy (LSC), small single-copy (SSC), and inverted repeat (IR) regions, respectively. The genes inside the circle were transcribed in a counterclockwise direction, while those outside the outer circle were transcribed in a clockwise direction. Different gene groups are represented by different colours.

The chloroplast genome of the newly assembled *R. fortunei* and 9 downloaded cp genomes were compared with the online mVISTA programme, and the annotation information of *R. fortunei* served as the reference in the Shuffle-LAGAN mode. Furthermore, multiple genome alignments were performed using MAUVE (v2.3.1) software for detecting rearrangements or inversions (Darling *et al.* 2004). The IR/LSC and IR/SSC junction regions were also compared with IRscope software to check whether expansion or contraction occurred in the *R. fortunei* cp genome (Amiryousefi *et al.* 2018).

Phylogenetic analysis of the Ericaceae cp genome. The phylogeny tree was constructed based on the complete chloroplast genome sequences of 2 species belonging to the *Gaultheria* genus (*G. fragrantissima* and *G. griffithiana*), 12 species of the *Rhododendron* species, 2 members of the *Vaccinium* genus (*V. oldhamii* and *V. macrocarpon*), and 6 species of the Ericaceae family (*Monotropa hypopitys*, *Pityopus californicus*, *Arbutus unedo*, *Hemitomes congestum*, *Pyrola rotundifolia*, and *Allotropa virgata*). These cp genomes were initially aligned using MAFFT v7 software with default settings for phylogenetic analysis, visualised by BioEdit, and evaluated by

Table 1 The genes identified in the *Rhododendron fortunei* chloroplast genome. The duplicated genes are in brackets.

Category of genes	Group of genes	Name of genes
Genes for photosynthesis	Subunits of ATP synthase	*atpA, atpB, atpE, atpF, atpH, atpI*
	Subunits of photosystem II	*psbA, psbB, psbC, psbD* (3×), *psbE, psbF, psbH, psbI, psbJ* (3×), *psbK, psbL, psbM, psbN, psbT, psbZ, ycf3*
	Subunits of NADH-dehydrogenase	*ndhA* (2×), *ndhB, ndhC, ndhD* (2×), *ndhE* (2×), *ndhF, ndhG* (2×), *ndhH* (2×), *ndhI* (2×), *ndhJ, ndhK*
	Subunits of cytochrome b/f complex	*petA* (2×), *petB, petD, petG, petL, petN*
	Subunits of photosystem I	*psaA, psaB, psaC* (2×), *psaI* (2×), *psaJ*
	Subunit of rubisco	*rbcL*
Self-replication	Large subunit of ribosome	*rpl14, rpl2, rpl20, rpl22, rpl32*(2×), *rpl33, rpl36*
	DNA dependent RNA polymerase	*rpoA, rpoB, rpoC1, rpoC2*
	Small subunit of ribosome	*rps11, rps14, rps16, rps18* (2×), *rps19* (3×), *rps2, rps3* (3×), *rps4, rps7, rps8, rps15* (3×)
Other genes	Subunit of Acetyl-CoA-carboxylase	*accD*
	c-type cytochrom synthesis gene	*ccsA* (2×)
	Envelop membrane protein	*cemA* (2×)
	Translational initiation factor	*infA*
	Maturase	*matK*
Unkown	Conserved open reading frames	*ycf15, ycf4* (2×), *nrgn*

IQ-TREE v2.0.3 (Yu *et al*. 2020). The phylogenetic trees were reconstructed and adjusted with RAxML v8.2.8 and Figtree v1.4 software, respectively (Alexandros 2014). RAxML v8.2.8 was used to run maximum likelihood (ML) analysis with a bootstrap value of 1000 (Alexandros 2014). The general time-reversible substitution model was used at normal settings in order to determine the rate of heterogeneity. The best models were selected using jModelTest v3.7 software (Posada 2008). Most importantly, the plastome of *Pyrola rotundifolia* (KU833271) was used as an outgroup.

RESULTS

General features of *R. fortunei* cp genomes. The Illumina paired-end run generated 21,133,214 paired end reads. After stringent quality assessment and filtering, 20,931,632 clean reads (3,136,355,677 clean bases) were obtained. The complete *R. fortunei* plastome sequence had a circular and quadripartite structure, with a total length of 200,998 bp and GC content of 41.23% (Fig.1). The large single-copy (LSC) region, small single-copy (SSC) region, and inverted repeat regions (IRs) were 109,151 bp (54.3%), 2,604 bp (1.3%), and 44,619 bp (22.19%), respectively (Fig. 1). Furthermore, the GC contents were 35.41%, 40.16%, and 29.46% in the LSC, SSC, and IR regions, respectively. Moreover, the Q20 (a base with a quality value greater than 20) and Q30 (a base with a quality value greater than 20) values were 97.19% and 92.11%, respectively.

In total, the *R. fortunei* plastome contained 147 genes, including 99 protein coding genes, 42 tRNA genes, and 6 rRNA genes. The lengths of the CDS, rRNA, tRNA, intergenic regions, and intron were 65,587 bp (32.63%), 8,808 bp (4.38%), 3,225 bp (1.6%), 46,995 bp (23.38%), and 76,680 bp (38.15%), respectively. Furthermore, the GC contents were 37.86%, 54.88%, 52.34%, 31.96%, and 33.71% in the CDS, rRNA, tRNA, intergenic regions, and intron region, respectively. Most of these genes were involved in photosynthesis, including subunits of ATP synthase, subunits of photosystem II, subunits of cytochrome b/f complex, subunits of photosystem I, subunits of NADH-dehydrogenase, and a subunit of Rubisco. In addition to self-replication, the c-type cytochrom synthesis gene, translational initiation factor, subunit of Acetyl-CoA-carboxylase, envelop membrane protein, and maturase were also found. In terms of photosynthesis, there were 6 subunits of ATP synthase (*atpA*, *atpB*, *atpE*, *atpF*, *atpH*, and *atpI*), 7 of photosystem I, 20 subunits of photosystem II, 17 of NADH-dehydrogenase, 7 subunits of cytochrome b/f complex, and 1 of Rubisco (*rbcL*) (Table 1). Regarding self-replication, 8 genes were large subunits of ribosome, 4 genes coded DNA-dependent RNA polymerase, and 18 genes were involved in the synthesis of small subunits of ribosome (Table 1). In addition, 7 genes were related to acetyl-CoA carboxylase, c-type cythochrome synthesis, envelope membrane protein, translational initiation factors, and maturase.

A total of 13 genes contained introns, consisting of *trnK-UUU*, *ycf3*, *trnL-UAA*, *trnC-ACA*, *rpoB*, *atpF*, *trnS-CGA*, *accD*, *rpl16*, *ndhB*, *trnE-UUC*, *ndhA*, and *trnA-UGC* (Table 2). With the exception of *ycf3,* all the other 12 genes had two exons and one intron, which ranged from 510 bp (*trnL-UAA*) to 104,936 bp (*rpl16*). For *ycf3*, the two introns were 711 bp and 742 bp, respectively. Exons I and II in *trnK-UUU* are 37 bp and 35 bp, and were separated by a 2,508-bp intron. Furthermore, exon I (9bp) and exon II (402 bp) are separated by a 104,936-bp intron in *rpl16*.

Table 2 The characteristics list of genes possessing intron.

Gene	Strand	Start	End	Exon I	Intron I	Exon II	Intron II	Exon III
trnK-UUU	-	1,841	4,420	37	2,508	35		
ycf3	-	6,865	8,824	124	711	232	742	151
trnL-UAA	+	11,475	12,069	35	510	50		
trnC-ACA	-	15,183	15,844	39	567	56		
rpoB	+	22,269	26,156	3,169	681	38		
atpF	+	36,008	37,275	161	701	406		
trnS-CGA	-	39,307	40,063	31	666	60		
accD	+	58,474	60,069	450	555	591		
rpl16	-	62,438	167,784	9	104,936	402		
ndhB	-	103,770	105,933	721	685	758		
trnA-UGC	+	116,189	117,068	37	807	36		
trnA-UGC	-	193,081	193,960	37	807	36		
ndhA	+	128,696	130,880	563	1,081	541		
ndhA	-	179,269	181,453	563	1,081	541		
trnE-UUC	+	115,111	116,124	32	942	40		
trnE-UUC	-	194,025	195,038	32	942	40		

Table 3 The relative synonymous codon usage in the *Rhododendron fortunei* plastome genome.

Amino acid	Codon	Number of occurrences	RSCU	Codon frequency per amino acid (%)	Amino acid	Codon	Number of occurrences	RSCU	Codon frequency per amino acid (%)
Ala	GCA	716	1.17	29.33	Pro	CCA	509	1.22	30.44
	GCC	352	0.58	14.42		CCC	301	0.72	18
	GCG	279	0.46	11.43		CCG	194	0.46	11.6
	GCU	1094	1.79	44.81		CCU	668	1.6	39.95
Cys	UGC	113	0.48	24.2	Gln	CAA	1072	1.58	78.99
	UGU	354	1.52	75.82		CAG	285	0.42	21
Asp	GAC	275	0.4	20.06	Arg	AGA	640	1.62	27.07
	GAU	1096	1.6	79.94		AGG	184	0.47	7.78
Glu	GAA	1434	1.54	77.06		CGA	609	1.55	25.76
	GAG	427	0.46	22.95		CGC	148	0.38	6.26
Phe	UUC	770	0.65	32.35		CGG	152	0.39	6.43
	UUU	1610	1.35	67.65		CGU	631	1.6	26.69
Gly	GGA	1080	1.5	37.49	Ser	AGC	192	0.41	6.84
	GGC	320	0.44	11.11		AGU	561	1.2	20
	GGG	482	0.67	16.73		UCA	501	1.07	17.86
	GGU	999	1.39	34.68		UCC	404	0.86	14.4
His	CAC	237	0.49	24.66		UCG	247	0.53	8.81
	CAU	724	1.51	75.34		UCU	900	1.92	32.09
Ile	AUA	1093	0.92	30.74	Thr	ACA	619	1.18	29.53
	AUC	684	0.58	19.23		ACC	417	0.8	19.89
	AUU	1779	1.5	50.03		ACG	220	0.42	10.5
Lys	AAA	1568	1.53	76.52		ACU	840	1.6	40.07
	AAG	481	0.47	23.47	Val	GUA	848	1.44	35.98
Leu	CUA	549	0.78	13.05		GUC	301	0.51	12.77
	CUC	254	0.36	6.04		GUG	316	0.54	13.41
	CUG	253	0.36	6.02		GUU	892	1.51	37.85
	CUU	892	1.27	21.21	Trp	UGG	732	1	100.01
	UUA	1455	2.08	34.59	Tyr	UAC	309	0.41	20.53
	UUG	803	1.15	19.09		UAU	1196	1.59	79.47
Met	AUG	955	1	100	Stop*	UAA	125	1.24	41.25
Asn	AAC	370	0.45	22.36		UAG	79	0.78	26.07
	AAU	1285	1.55	77.65		UGA	99	0.98	32.68

Table 4 The frequency of microsatellite types in the *Rhododendron fortunei* cp genome.

Repeats	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	total	Percentage
A/T	-	-	-	-	-	33	19	5	5	3	2			2	2	71	0.922
AT/AT	-	4											1			5	0.065
AAT/ATT	1															1	0.013

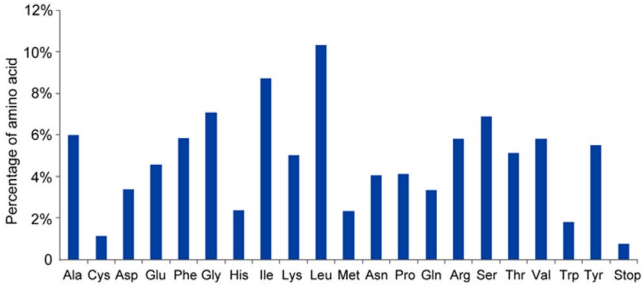

Fig. 2. The percentage of protein-coding amino acids in the *Rhododendron fortunei* chloroplast genome.

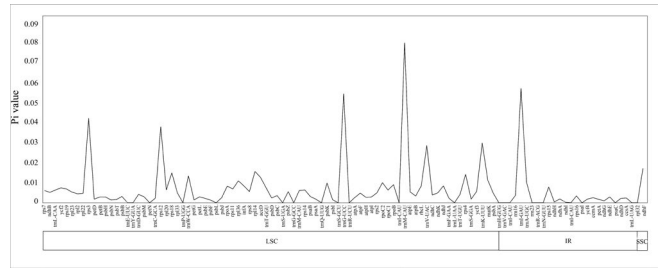

Fig. 3. The nucleotide diversity of 10 chloroplast genomes of the *Rhododendron* genus. The X-axis presents the position of the aligned chloroplast genomes, and the Y-axis refers to nucleotide diversity. Below the X-axis, the LSC, SSC, and IR regions are displayed with arrows.

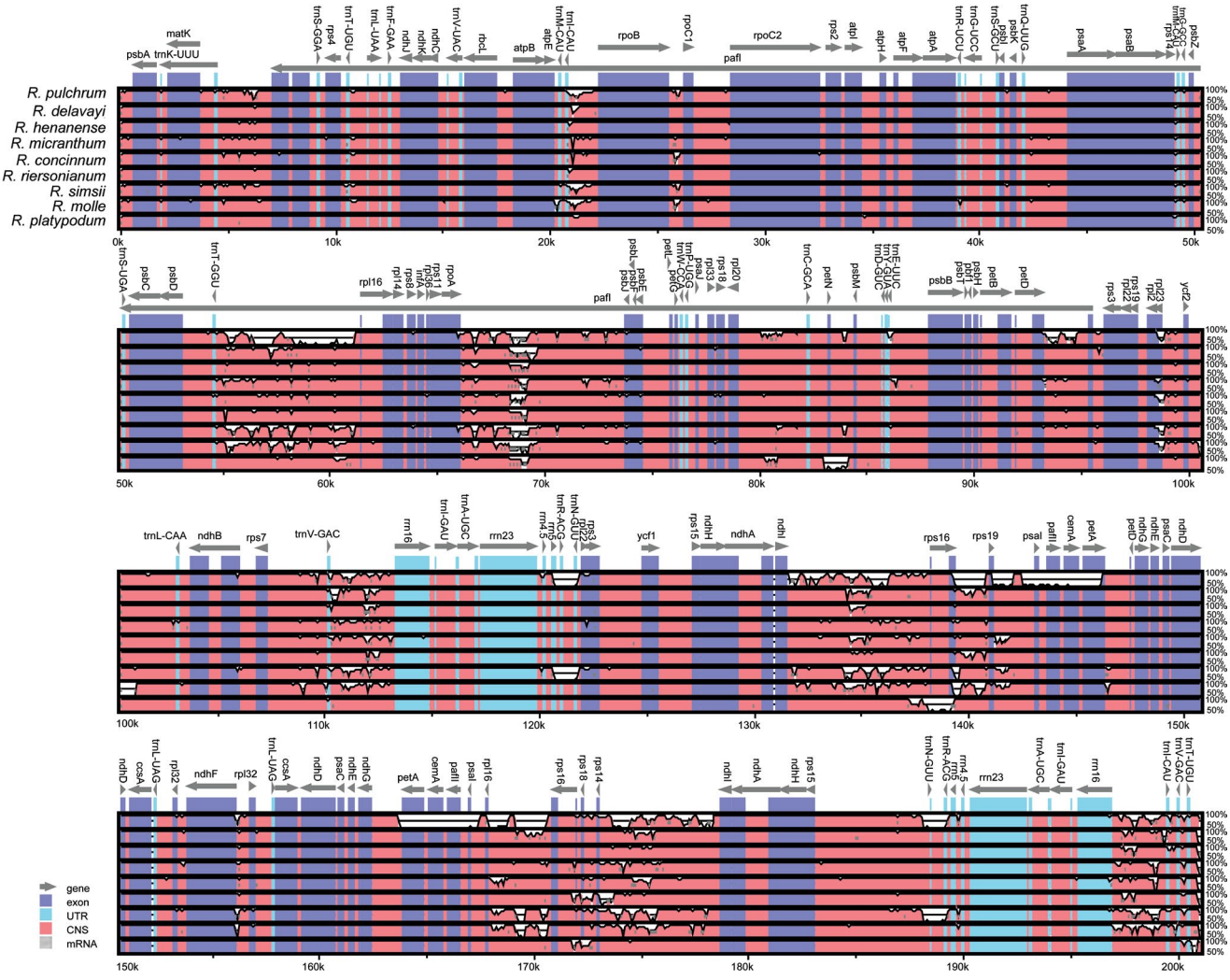

Fig. 4. The comparison of 10 cp genomes with *Rhododendron fortunei* annotation as a reference. These genome regions were colour-coded as exons, introns, and conserved non-coding sequences (CNS), respectively. The vertical scale represents the percentage of identity ranging from 50% to 100%. The horizontal axis presents the coordinates within the cp genome.

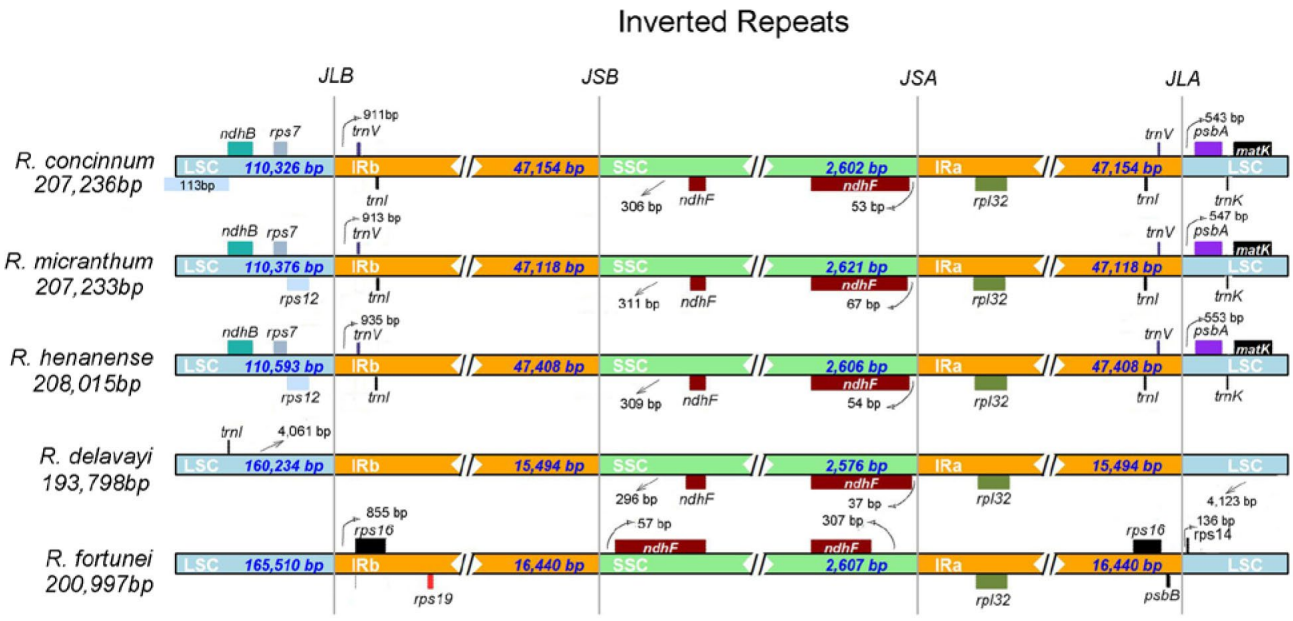

Fig. 5. The comparison of the LSC, SSC, and IR regional boundaries of the cp genome between *Rhododendron fortunei* and five related taxa. The JLB, JSB, JSA, and JLA present the "junction line between LSC and IRb", the "junction line between IRb and SSC", the "junction line between SSC and Ira", and the "junction line between IRa and LSC", respectively.

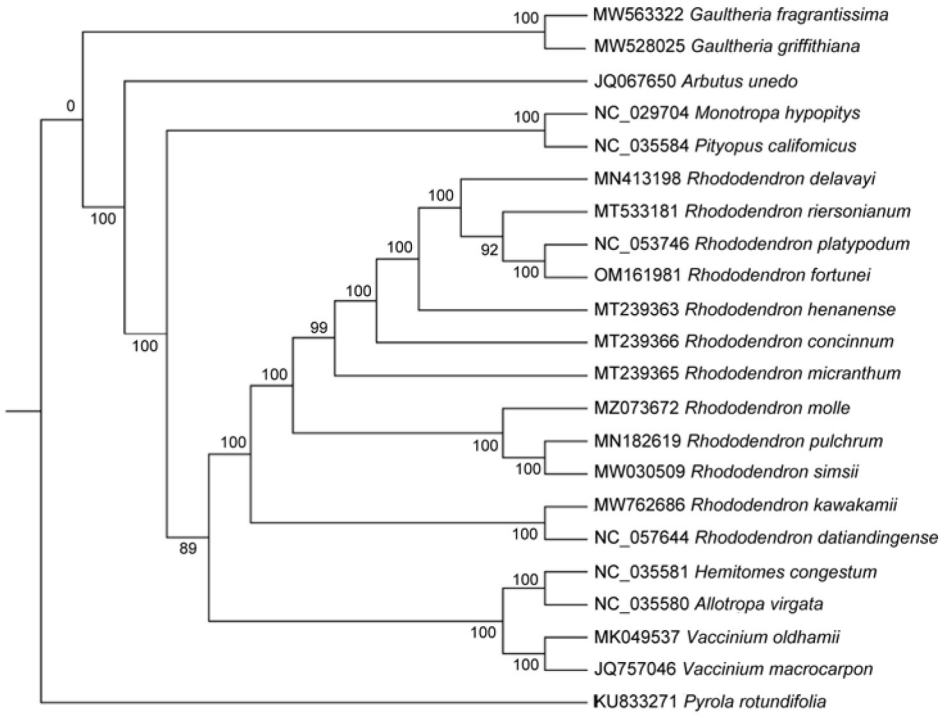

Fig. 6. The maximum-likelihood phylogenetic tree for *Rhododendron fortunei*. The numbers on each node refer to the bootstrap support values.

Codon usage and SSR analysis of the *R. fortunei* plastome. Protein coding nucleotides were used to compute the codon usage bias of the *R. fortunei* plastome. A total of 40,706 amino acid codons were found (Table 3). Leucine (Leu) was the most frequent amino acid (11.51%), followed by isoleucine (Ile) and glycine (Gly), accounting for 8.74% and 7.08%, respectively (Fig. 2), while cysteine (Cys) was the lowest representative amino acid (1.15%). Based on the RSCU values, 30 codons showed obvious codon usage bias, as the RSCU values were all greater than 1 (Table 3). All these codons with usage bias were A/U-ending codons. For codons with RSCU values of less than 1, C/G-ending codons were predominant.

Microsatellites, a group of short repeat sequences (1–6 bp), are tools used to assess molecular diversity and reveal genetic variation (Kaila *et al*. 2017). In this study, 77 SSRs were identified from the *R. fortunei* cp genome, and 9 were present in compound formation. In our study only mononucleotide (71), dinucleotide (5), and trinucleotide (1) motifs were found, accounting for 92.208%, 6.494%, and 1.299%, respectively. All these mononucleotide motifs were A/T repeats, and those with repeat numbers of 10-15 were the most abundant. Furthermore, four (A/T)n microsatellites repeated more than 18 times were also detected (Table 4). In terms of the dinucleotide motifs, four AT/AT types with five repeat times and one (TA)17 were found, while for trinucleotide motifs, a (TAA)5 microsatellite was screened. The microsatellite density in the intergenic spacer regions was significantly higher (62.34%) than in the coding regions (37.66%). Among the 29 microsatellites distributed in the gene coding region, 22 were found in the *rpl16* gene, while the other 7 repeat motifs were detected in the genes *matK*, *ndhA*, *rpoA*, *rps7*, *rps8*, *ccsA*, and *cemA*.

Analysis of nucleotide diversity, plastome sequence divergence, and hotspot regions. Nucleotide diversity analysis was carried out to investigate the divergence levels between different cp genomes of *Rhododendron* species. The Pi values ranged from 0 to 0.007977. In particular, a high average level of genetic variation occurred in the SSC region (Pi=0.01712), followed by the LSC region (Pi=0.00733) and IR region (Pi=0.00249) (Fig. 3). Relatively high levels of nucleotide diversity (Pi>0.03) were found in 5 gene regions, containing *trnM-CAU* (Pi=0.07977), *trnI-GAU* (Pi=0.05709), *trnG-UCC* (Pi=0.05429), *rps3* (Pi=0.0422), and *rps12* (Pi=0.03788). Furthermore, *trnK-UUU* (Pi=0.02982) and *trnV-UAC* (Pi=0.02857) also exhibited high Pi values, showing great potential for the development of species-specific markers.

The structural characteristics of the Ericaceae cp genomes were investigated using mVISTA software, with the annotated *R. fortunei* cp genome serving as a reference. The alignment outcome revealed highly conserved genomes with a few variations, and the coding regions were more conserved than the non-coding regions (CNS in Fig. 4), which is the same as other flowering plants. The LSC regions proved to be more stable than the IR regions. Seven highly divergent intergenic regions were found, containing *trnI-rpoB*, *trnT-rpl16*, *rpoA-psbJ*, *rps7-rrn16*, *ndhI-rps16*, *rps16-rps19*, and *rrn16-trnI*. However, slight variations were also observed in the genes *trnK-UUU*, *psbJ*, *trnR-ACG*, *trnN-GUU*, *rpl23*, *trnR-ACG*, *rpl22*, *rps16*, *trnI-CAU*, and *rrn16* due to intron regions. MAUVE analysis showed no rearrangements or inversions in the *R. fortunei* cp genome.

The IR regions of these 5 *Rhododendron* species ranged from 15,494 bp (the *R. fortunei* cp genome) to 47,408 bp (the *R. henanense* cp genome). Some expansion and contraction existed in the IR regions. In the *R. concinnum*, *R. micranthum*, and *R. henanense* cp genomes, the line between LSC and IRb (JLB line) was located between genes *rps7* and *trnV*, while *trnV* was located in the IRb region with a length of 911-935 bp extending to the LSC region (Fig. 5). However, the *trnI* gene was located in the LSC region with a 4,061 bp extending to the IRb region in the *R. delavayi* cp genome. In the *R. fortunei* cp genome, the JLB line was located near gene *rps16*, which was located in the IRb region with 855 bp extending to the LSC region (Fig. 5). In the *R. concinnum*, *R. micranthum*, *R. henanense*, *R. delavayi*, and *R. fortunei* cp genomes, the *ndhF* gene was located in the SSC region with extending regions of 306 bp, 311 bp, 309 bp, 296 bp, and 57 bp to the JSB line, respectively. Meanwhile, another *ndhF* gene *was* also located in the SSC region with 53 bp, 67 bp, 54 bp, 37 bp, and 307 bp to the JSA line in the *R. concinnum*, *R. micranthum*, *R. henanense*, *R. delavayi*, and *R. fortunei* cp genomes, respectively. In the *R. concinnum* and *R. micranthum* cp genomes, genes *trnV* and *psbA* were located at the junction of the IRa/LSC region, while gene *psbA* was located in the LSC region with 543 bp-553 bp extending to the IRa region. However, genes *rps16* and *rps14* were located at the junction of the IRa/LSC region, while *rps14* was located in the LSC region with 136 bp extending to the IRa region in the *R. fortunei* cp genome.

Phylogenetic analysis. In order to clarify the phylogenetic location of *R. fortunei* among the Ericaceae family, the complete *R. fortunei* cp genome and a further five species cp genomes belonging to the Ericaceae family were used to reconstruct the phylogenetic relationships. Based on the phylogenetic tree, all these 22 species belonging to the Ericaceae family were grouped into one clade and clustered into two subclades. In the phylogenetic tree, the nodes showed bootstrap values greater than 89%. *R. fortunei* was most closely related to *R. platypodum*, and highly related to *R. riersonianum* and *R. delavayi* (Fig. 6). Furthermore, all these taxa belonging to the *Rhododendron* genus were grouped together. Compared with *M. hypopitys* and *P. californicus,* the cp genomes of *H. congestum*, *A. virgata*, *V. oldhamii*, and *V. macrocarpon* were more closely related to that of *R. fortunei*. The topological structure was almost consistent with the previously published phylogeny (Liu *et al*. 2021).

DISCUSSION

The plant plastome serves as a good model for investigating lineage-specific molecular evolution, and is valuable in comparative genomic research and phylogenomic analyses due to the polymorphic regions generated through genomic expansion, contraction, inversion, and even gene rearrangement (Sanitá Lima *et al*. 2016; Cauz-Santos *et al*. 2020; Wang *et al*. 2020).

As the main organelle in plants, chloroplast genes play critical roles in transforming light energy into chemical energy, which also undergo adaptive evolution (Zhang *et al*. 2018a). However, the availability of cp genome information is relatively scarce in the Ericaceae family. In addition, species of the Ericaceae family are relatively difficult to distinguish based on morphological and chemical data. This study presented the complete chloroplast genome of *R. fortunei* in order to evaluate the evolutionary relationships among the Ericaceae family. The *R. fortunei* cp genome (200,998 bp) is of a typical quadripartite structure. However, no inverted repeat regions existed in the cp genome of *R. pulchrum*, with a length of 136,249 bp (Shen *et al*. 2020). Moreover, the GC content of the *R. fortunei* cp genome was (41.23%), which was larger than that of *Myracrodruon urundeuva* (37.8%), *R. pulchrum* (35.98%) and the Rubus species (37.0%–37.3%) (Shen *et al*. 2020; Rossini *et al*. 2021; Yu *et al*. 2022).

Like other higher plants, the 99 protein-coding genes were mainly involved in self-replication and photosynthesis (Yu *et al*. 2022). In total, 13 genes contained introns, and the highly variable introns might be the main reason for the cp genome size (Park *et al*. 2017). The ycf1 and ycf2 genes, two of the largest open reading frames in angiosperms, were absent in the *R. fortunei* chloroplast genome. In the cp genome of *Common bermudagrass*, the ycf1, ycf2, ycf15, and ycf68 genes are pseudogenised (Huang *et al*. 2017). It is possible that the functional ycf1 and ycf2 genes might be transferred to the nuclei, similar to the accD gene in several species of Poaceae (Huang *et al*. 2017). Most of the protein coding genes of the *R. fortunei* cp genome start with a typical ATG codon (coding methionine), while others begin with codons ATC, GTG and ACG, which are the same as most angiosperm plant chloroplast genomes (Raman & Park 2016; Li *et al*. 2017). Codon usage can greatly affect the chloroplast genome evolution, and evolutionary phenomena are the result of mutation bias (Li *et al*. 2017). Leucine is the most frequent codon in the *R. fortunei* plastome, which is similar to other flowering plant genomes, such as *M. urundeuva* (Liu *et al*. 2018; Rossini *et al*. 2021). Codon bias is an efficient translation mechanism influenced by mutation pressure and natural selection (Zhang *et al*. 2022). Like *M. urundeuva* and Solanum (Zhang *et al*. 2018b; Rossini *et al*. 2021), 30 codon usage biases (values > 1) were observed for A/U-ending codons in the *R. fortunei* cp genome.

Molecular markers developed from the cp genome, such as plastid genes *rbcL*, *psbA* and the nuclear internal transcribed spacer (ITS), have played a significant role in species identification (Trobajo *et al*. 2010; Liu *et al*. 2011). For example, the cp SSRs contributed a great deal to the genetic improvement in pears (Yue *et al*. 2018). In total, 77 SSRs were identified from the *R. fortunei* cp genome, containing only mononucleotide, di-nucleotide, and tri-nucleotide repeats. Furthermore, high richness in mononucleotide repeats (A/T)n has been observed, a characteristic shared by the cp genomes of most flowering plants (Jo *et al*. 2017; Li *et al*. 2017; Santos & Almeida 2019; Rossini *et al*. 2021). The number of microsatellites identified in the *R. fortunei* cp genome was slightly higher than that of *Mangifera indica* (57 SSRs) and *Spondias bahiensis* (53 SSRs), but lower than that of *Syringa pinnatifolias* (253 SSRs) (Jo *et al*. 2017; Santos & Almeida 2019). Furthermore, no tandem guanine (G) and cytosine (C) repeats were identified in the *R. fortunei* cp genome, as high rates of A/T may be closely related to the high content of A and T bases.

According to neutral theory, nucleotide substitutions in the intergenic spacer, intron region, and pseudogenes are considered to be almost neutral or near-neutral, making it unlikely for them to have been affected by natural selection (Akashi *et al*. 2012). Therefore, the molecular evolution which occurred in the non-coding region could provide valuable insights into the evolutionary history of *R. fortunei*. Compared with the coding regions, the non-coding regions usually mutate more rapidly (Yu *et al*. 2022). These regions, containing *trnI-rpoB*, *trnT-rpl16*, *rpoA-psbJ*, *rps7-rrn16*, *ndhI-rps16*, *rps16-rps19*, and *rrn16-trnI*, could serve as the first candidates for developing molecular markers to indentify *R. fortunei* species. A high degree of similarity was detected among these tested Ericaceae species, and the coding regions were more conserved than the non-coding regions. Notably, the LSC regions were more stable than the IR regions.

Gene changes occurring during selection stress could lead to the rapid transformation of genes into new adaptive combinations, and help plants to adapt to new habitats (Xie *et al*. 2018). The expansion and contraction of the IR region boundaries have been identified as the main drivers of size changes in the cp genome, thus playing an important role in species evolution (Yang & Dos Reis 2010). In the plastid genome of *R. delavayi* (202,169 bp) with abundant repeat sequences, the rearrangement and inversion occurred mainly in the large single copy region, while the extreme expansion of the inverted repeat region served to shorten the small single copy region, while expanding the full length of the genome (Li *et al*. 2020). Therefore, these genetic variations present in the *R. fortunei* cp genome are believed to facilitate the adaptation of *R. fortunei* to changing survival conditions. The high variability in both the coding and non-coding regions, such as the *trnI-rpoB*, *trnT-rpl16*, and *rpoA-psbJ* regions, will provide a solid foundation for phylogenetic analysis and species identification in the Ericaceae family. *R. fortunei* and *R. platypodum* were firstly clustered, and then grouped with *R. riersonianum* and *R. delavayi*, inferring that these *Rhododendron* species shared a similar genetic structure and pressure adaptability. Furthermore, the topological structure was almost consistent with the previously published phylogeny of Ericaceae species (Liu *et al*. 2021). The whole

plastome will serve as a reliable marker for phylogenetic research of the Ericaceae family.

Data availability statement
The cp genome of *R. fortunei* was submitted to GenBank under accession number OM161980.

Acknowledgment – This work was supported by the Scientific and Technological Research Project of the Hubei Provincial Department of Education (D20222902), and the Open fund of Hubei Key Laboratory of Economic Forest Germplasm Improvement and Resources Comprehensive Utilization (202303202 and 202140804).

REFERENCES

Abdullah Henriquez C, Mehmood F, Hayat A & Ahmed I. 2021. Chloroplast genome evolution in the *Dracunculus* clade (Aroideae, Araceae). *Genomics* **113**: 183–192.

Akashi H, Osada N & Ohta T. 2012. Weak selection and protein evolution. *Genetics* **192**: 15–31.

Alexandros S. 2014. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.

Amiryousefi A, Hyvonen J & Poczai P. 2018. IRscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics* **34**: 3030–3031.

Cauz-Santos LA, Da Costa ZP, Callot C, Cauet S, Zucchi MI, Bergès H, van Den Berg C & Vieira MLC. 2020. A repertory of rearrangements and the loss of an inverted repeat region in passiflora chloroplast genomes. *Genome Biology Evolution* **12**: 1841–1857.

Daniell H, Lin CS, Yu M & Chang WJ. 2016. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biology* **17**: 134.

Darling ACE, Mau B, Blattner FR & Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research* **14**: 1394–1403.

Dong W, Xu C, Cheng T, Lin K & Zhou SL. 2013. Sequencing angiosperm plastid genomes made easy: a complete set of universal primers and a case study on the phylogeny of Saxifragales. *Genome Biology Evolution* **5**: 989–997.

Dong W, Xu C, Li C, Sun J, Zuo Y, Shi S, Cheng T, Guo J & Zhou S. 2015. ycf1, the most promising plastid DNA barcode of land plants. *Scientific Reports* **5**: 8348.

Givnish TJ, Zuluaga A, Spalink D, Soto Gomez M, Lam VKY, Saarela JM, Sass C, Iles WJD, De Sousa DJL, Leebens-Mack J, Chris PJ, Zomlefer WB, Gandolfo MA, Davis JI, Stevenson DW, De Pamphilis C, Specht CD, Graham SW, Barrett CF & An C. 2018. Monocot plastid phylogenomics, timeline, net rates of species diversification, the power of multi-gene analyses, and a functional model for the origin of monocots. *American Journal of Botany* **105**: 1888–1910.

Greiner S, Lehwark P & Bock R. 2019. Organellar genome DRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Research* **47**: W59–W64.

Hanussek M, Bartusch F & Krüger J. 2021. Performance and scaling behavior of bioinformatic applications in virtualization environments to create awareness for the efficient use of compute resources. *PLoS Computational Biology* **17**: e1009244.

Huang YY, Cho ST, Haryono M & Kuo CH. 2017. Complete chloroplast genome sequence of common bermudagrass (*Cynodon dactylon* (L.) Pers.) and comparative analysis within the family poaceae. *PLoS One* **12**(6): e0179055.

Jo S, Kim HW, Kim YK, Sohn JY, Cheon SH & Kim KJ. 2017. The complete plastome sequences of *Mangifera indica* L. (Anacardiaceae). *Mitochondrial DNA B* **2**: 698–700.

Kaila T, Chaduvla PK, Rawal HC, Saxena S, Tyagi A, Mithra SV, Solanke AU, Kalia P, Sharma TR, Singh NK & Gaikwad K. 2017. Chloroplast genome sequence of clusterbean (*Cyamopsis tetragonoloba* L.): Genome structure and comparative analysis. *Genes* **8**: 212.

Li B, Lin F, Huang P, Guo W & Zheng Y. 2017. Complete chloroplast genome sequence of *Decaisnea insignis*: Genome organization, genomic resources and comparative analysis. *Scientific Reports* **7**: 10073.

Li H, Guo Q, Li Q & Yang L. 2020. Long-reads reveal that *Rhododendron delavayi* plastid genome contains extensive repeat sequences, and recombination exists among plastid genomes of photosynthetic Ericaceae. *PeerJ* **8**: e9048.

Liu C, Liang D, Gao T, Pang X, Song J, Yao H, Han J, Liu Z, Guan X, Jiang K, Li H & Chen S. 2011. PTIGS-IdIt, a system for species identification by DNA sequences of the psbA-trnH intergenic spacer region. *BMC Bioinformatics* **12**: S4.

Liu C, Shi L, Zhu Y, Chen H, Zhang J, Lin X & Guan X. 2012. Cpgavas, an integrated web server for the annotation, visualization, analysis, and genbank submission of completely sequenced chloroplast genome sequences. *BMC Genomics* **13**: 715.

Liu HJ, Ding CH, He J, Cheng J, Pei LY & Xie L. 2018. Complete chloroplast genomes of *Archiclematis*, *Naravelia* and *Clematis* (Ranunculaceae), and their phylogenetic implications. *Phytotaxa* **343**: 214–226.

Liu Y, Li Q, Wang L, Wu L, Huang Y, Zhang J, Song Y & Liao J. 2021. The complete chloroplast genome of *Rhododendron molle* and its phylogenetic position within Ericaceae. *Mitochondrial DNA Part B* **6**: 2587–2588.

Luo J, Hou BW, Niu ZT, Liu W, Xue QY & Ding XY. 2014. Comparative chloroplast genomes of photosynthetic orchids: insights into evolution of the Orchidaceae and development of molecular markers for phylogenetic capplications. *PLoS One* **9**: e99016.

Morton BR. 2022. Context-dependent mutation dynamics, not selection, explains the codon usage bias of most angiosperm chloroplast genes. *Journal of Molecular Evolution* **90**: 17–29.

Park I, Kim WJ, Yeo SM, Choi G, Kang YM, Piao R & Moon BC. 2017. The complete chloroplast genome sequences of *Fritillaria ussuriensis* Maxim. and comparative analysis with other *Fritillaria* species. *Molecules* **22**: 982.

Popescu R & Kopp B. 2013. The genus *Rhododendron*: An ethnopharmacological and toxicological review. *Journal of Ethnopharmacology* **147**: 42–62.

Posada D. 2008. jModelTest: Phylogenetic model averaging. *Molecular Biology Evolution* **25**: 1253–1259.

Raman G & Park S. 2016. The complete chloroplast genome sequence of *Ampelopsis*: Gene organization, comparative analysis, and phylogenetic relationships to other angiosperms. *Frontiers in Plant Science* **7**: 341.

Rossini BC, Moraes MD & Marino CL. 2021. Complete chloroplast genome of *Myracrodruon urundeuva* and its phylogenetics relationships in Anacardiaceae family. *Physiology and Molecular Biology of Plants* **27**: 801–814.

Sanitá Lima M, Woods LC, Cartwright MW & Smith DR. 2016. The (in)complete organelle genome: exploring the use and nonuse of available technologies for characterizing mitochondrial and plastid chromosomes. *Molecular Ecology Resources* **16**: 1279–1286.

Santos V & Almeida C. 2019. The complete chloroplast genome sequences of three *Spondias* species reveal close relationship among the species. *Genetics and Molecular Biology* **42**: 132–138.

Schmieder R & Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**: 863–864.

Shen J, Li X, Zhu X, Huang X & Jin S. 2020. The complete plastid genome of *Rhododendron pulchrum* and comparative genetic analysis of Ericaceae species. *Forests* **11**(2): 158.

Shi W, Song W, Chen Z, Cai H, Gong Q, Liu J, Shi C & Wang S. 2023. Comparative chloroplast genome analyses of diverse *Phoebe* (Lauraceae) species endemic to China provide insight into their phylogeographical origin. *PeerJ* **11**: e14573.

Smith DR. 2015. Mutation rates in plastid genomes: they are lower than you might think. *Genome Biology Evolution* **7**: 1227–1234.

Thiel T, Michalek W, Varshney R & Graner A. 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theoretical and Applied. Genetics* **106**: 411–422.

Trobajo R, Mann DG, Clavero E, Evans KM, Vanormelingen P & McGregor RC. 2010. The use of partial *cox*1, *rbc*L and LSU rDNA sequences for phylogenetics and species identification within the *Nitzschia palea* species complex (Bacillariophyceae). *European Journal of Phycology* **45**: 413–425.

Wambugu PW, Brozynska M, Furtado A, Waters DL & Henry RJ. 2015. Relationships of wild and domesticated rices (*Oryza* AA genome species) based upon whole chloroplast genome sequences. *Scientific Reports* **5**: 13957.

Wang S, Jin Z, Luo Y, Li Z, Fang Y, Xiang J & Jin W. 2019. Genetic diversity and population structure of *Rhododendron simsii* (Ericaceae) as reveled by microsatellite markers. *Nordic Journal of Botany* **37**: 1–10.

Wang S, Li Z, Guo X, Fang Y, Xiang J & Jin W. 2018. Comparative analysis of microsatellite, SNP, and Indel markers in four *Rhododendron* species based on RNA-seq. *Breeding Science* **68**: 536–544.

Wang X, Rhein HS, Jenkins J, Schmutz J & Randall JJ. 2020. Chloroplast genome sequences of *Carya illinoinensis* from two distinct geographic populations. *Tree Genetics & Genomes* **16**: 48.

Xie DF, Yu Y, Deng YQ, Li J, Liu HY, Zhou SD & He XJ. 2018. Comparative analysis of the chloroplast genomes of the Chinese endemic genus *Urophysa* and their contribution to chloroplast phylogeny and adaptive evolution. *International Journal of Molecular Sciences* **19**: 1847.

Yang Z & Dos Reis M. 2010. Statistical properties of the branch-site test of positive selection. *Molecular Biology Evolution* **28**: 1217–1228.

Yap JY, Rohner T, Greenfield A, Van Der Merwe M, McPherson H, Glenn W, Kornfeld G, Marendy E, Pan AY, Wilton A, Wilkins MR, Rossetto M & Delaney SK. 2015. Complete chloroplast genome of the Wollemi pine (*Wollemia nobilis*): structure and evolution. *PLoS One* **106**: e0128126

Yu J, Dong H, Xiang J, Xiao Y & Fang Y. 2020. Complete chloroplast genome sequence and phylogenetic analysis of *Magnolia pilocarpa*, a highly ornamental species endemic in central China. *Mitochondrial DNA B* **5**: 720–722.

Yu JJ, Fu J, Fang YP, Xiang J & Dong HJ. 2022. Complete chloroplast genomes of *Rubus* species (Rosaceae) and comparative analysis within the genus. *BMC Genomics* **23**: 32.

Yue X, Zheng X, Zong Y, Jiang S, Hu C, Yu P, Liu G, Cao Y, Hu H & Teng Y. 2018. Combined analyses of chloroplast DNA haplotypes and microsatellite markers reveal new insights into the origin and dissemination route of cultivated pears native to East Asia. *Frontiers in Plant Science* **9**: 591.

Zhang J, Huang H, Qu C, Meng X & Xing S. 2022. Comprehensive analysis of chloroplast genome of *Albizia julibrissin* Durazz. (Leguminosae sp.). *Planta* **255**: 26.

Zhang R, Zhang L, Wang W & Zhang Z. 2018b. Differences in codon usage bias between photosynthesis-related genes and genetic system-related genes of chloroplast genomes in cultivated and wild solanum species. *International Journal of Molecular Sciences* **19**: 3142.

Zhang Y, Iaffaldano BJ, Zhuang X, Cardina J & Cornish K. 2017. Chloroplast genome resources and molecular markers differentiate rubber dandelion species from weedy relatives. *BMC Plant Biology* **17**: 34.

Zhang ZH, An ML, Miao JL, Gu ZQ, Liu C & Zhong BJ. 2018a. The Antarctic sea ice alga *Chlamydomonas* sp. ICE-L provides insights into adaptive patterns of chloroplast evolution. *BMC Plant Biology* **18**: 53.

REZIME

Botanica SERBICA

Kompletna sekvenca hloroplastnog genoma *Rhododendron fortunei*: strukturna komparativna i filogenetička analiza u familiji Ericaceae

Yunli Xiao, Wanjing Zhang, Yirong Sun, Zhiliang Li, Jiaojun Yu, Chunyu Zhang i Shuzhen Wang

Rhododendron fortunei (Ericaceae) je vrsta sa hortikulturnim i lekovitim vrednostima. Međutim, informacije o genomu ove vrste su veoma ograničene. U ovoj studiji je sastavljen i zabeležen kompletan hloroplastni genom (cp) *R. fortunei*, okarakterisani SSR lokusi, urađena komparativna genomička analiza, kao i filogenetska istraživanja. Rezultati su pokazali da je hloroplastni genom *R. fortunei* sa tipičnom kvadripartitnom strukturom (200,997 bp). Konkretno, dužine regiona velike pojedinačne kopije (LSC), regiona obrnutih ponavljanja (IR) i regiona male pojedinačne kopije (SSC) bile su 109,151 bp, 2,604 bp i 44,619 bp, respektivno. Identifikovano je ukupno 147 jedinstvenih gena, uključujući 99 gena koji kodiraju proteine, 42 gena tRNA i 6 rRNA gena. Leucin (11,51%) i cistein (1,15%) su bile najviše i najmanje zastupljene aminokiseline, respektivno. Konkretno, 30 kodona sa očiglednom sklonošću upotrebe kodona bili su svi sa A/U-završetkom. Među 77 ponavljanja jednostavne sekvence, većina su bila mononukleotidna A/T ponavljanja smeštena u intergenskom razmaku. Pet genskih regiona pokazalo je visok nivo nukleotidne raznovrsnosti (Pi > 0,03). Komparativna analiza genoma otkrila je 7 intergenskih regiona žarišta (*trnI-rpoB, trnT-rpl16, rpoA-psbJ, rps7-rrn16, ndhI-rps16, rps16-rps19* i *rrn16-trnI*), pokazujući velike potencijale za stvaranje molekula za vrste. Ekspanzija i kontrakcija su detektovane u IR regionu *R. fortunei* cp genoma. U filogenetskom stablu, *R. fortunei* je blisko povezan sa *R. platipodum*. Ovo istraživanje će biti značajno za proučavanje evolucionog i genetskog diverziteta *R. fortunei* i srodnih vrsta iz porodice Ericaceae.

Ključne reči: *Rhododendron fortunei*, sekvenciranje sledeće generacije, hloroplastni genom, komparativna genomika, konzervaciona genetika