The chloroplast genome sequence and phylogenetic analysis of Rubia alata Wall and Rubia ovatifolia Z. Ying Zhang.(Rubiaceae)

JiaZhou Shi
University of Chinese Medicine
XiaoYing Chen
University of Chinese Medicine
YiYao Jing
University of Chinese Medicine
Yonggang Yan
University of Chinese Medicine
Gang Zhang
University of Chinese Medicine
BingYue Yang
University of Chinese Medicine
Liang Peng (ppengliang@126.com)
University of Chinese Medicine

Research Article

Keywords: Chloroplast genome, Rubia alata Wall, Rubia ovatifolia Z. Ying Zhang, Phylogenetic analysis

Posted Date: February 5th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3896731/v1

Additional Declarations: No competing interests reported.

Abstract

Background Rubiaceae, a genus encompassing approximately 70 species distributed globally, includes Rubia alata Wall and Rubia ovatifolia Z. Ying Zhang, distinctive medicinal plants native to China. The acquisition of chloroplast genome sequences of *R. alata* and *R. ovatifolia* holds paramount significance for elucidating species divergence, constructing phylogenetic relationships, and comprehending RNA editing processes

Result In this study, the complete chloroplast genome of *R. alata* and *R. ovatifolia* were obtained based on the sequences generated by Illumina HiSeq. The chloroplast genome of *R. alata* is 154,973 base pairs (bp) in length, containing a large single-copy region (LSC) of 84,801 bp, a small single-copy region (SSC) 17,138 bp, and a pair of inverted repeats (IRs) of the same length. The length is 26,517 bp, and the length of the LSC, SSC, and IR regions of *R. ovatifolia* is 84,716, 17,116 and 26,517 bp, respectively. The chloroplast genomes in both varieties cover 130 genes, including 85 protein-coding genes, 37 transfer-RNA genes (tRNA), 8 ribosomal RNA genes (rRNA). There were differences in the size of the chloroplast genome, while the genome structure was conservative and the composition was similar. Although there was a small degree of variation, it still had high homology. The assembled chloroplast genome in this study provided useful molecular information about the evolution of *R. alata* and *R. ovatifolia* which was a basis for phylogenetic analyses of Rubiaceae.

Conclusions The findings of this study have improved our understanding of the internal structure of the chloroplast genome of Rubia plants and are of great value for the future development and utilization of *R. alata* and *R. ovatifolia*resources.

1. Introduction

About 70 species of the genus Rubia were found in all over the world; 36 of those species and two variants were reported in China[]. Rubia species, one of the earliest plant resources have important commercial and medicinal values. In the past, these species were utilized commercially as natural dyestuffs that enhanced the movement of goods; in the medical field, their use as medications was initially documented in the Divine Famer's Materia Medica, the world-famous pharmacy book of China with a history spanning over two millennia[]. Numerous medical literature claim that Chinese traditional medicine widely employed the roots of rubia plants, which were said to have satisfactory efficacy, to cure a variety of conditions, including cancers, tuberculosis, rheumatism, hematemesis, metrorrhagia, epistaxis, contusion and menoxenia[,].

Rubia alata Wall, an endemic species in China is a herbaceous climbing vine that thrives in the Yangtze River Basin and adjacent southern regions[]. *R. alata* is bitter, astringent and cool. It is mainly used to treat heat stroke, fever, sore throat, chest pain, hemoptysis, nephritis, jaundice, diarrhea, dysentery, rheumatic joint pain, infantile malnutrition, hookworm disease, scabies and other diseases. The whole grass can be used as herbal tea boiled water to drink, and can also be used as medicinal materials and meat stew

soup to eat *R. alata* has a cold, astringent, and bitter flavor. It is primarily used to treat a variety of illnesses, including scabies, infantile malnutrition, rheumatic joint pain, hemoptysis, nephritis, fever, sore throat, chest pain, and diarrhea. The entire grass can be used to make herbal tea, boil water, and make medicinal materials as well as beef stew soup. Two novel naphthohydroquinone dimers have currently been identified from the medicinal plant *R. alata* and its precursor rubiadin, and their cytotoxicity of 1–3 as well as their impact on the NF-κB pathway have been assessed, according to research[], Another herbaceous climbing plant in the Rubia genus is *Rubia ovatifolia* Z. Ying Zhang, which is produced in Shaanxi, Gansu, Zhejiang, and other regions of China and is likewise endemic. The habitat restricts between 1700 and 2200 meters high, in mountainous areas with few bushes or trees. Given the herb's medicinal and therapeutic properties, research on its heredity and variation is extremely important[].

Plants have self-replicating organelles called chloroplasts, which are made up of their own genomes that encode proteins necessary for photosynthesis and other biological functions[,]. The chloroplast genomes exhibit a highly conserved structure with a typical quadripartite structure that includes a pair of inverted repeats (IRs), separated by a large single-copy (LSC) region and a small single-copy (SSC) region among seed plants[,,]. Apart from the preservation of the chloroplast genome's structure, variations exist between species and even within individuals, such as gene loss, rearrangements in sequence, indels, and the expansion or shrinkage of the IR, which offers great values for plant evolution and taxonomic analysis[,,]. Furthermore, the majority of angiosperms have haploid chloroplast genomes, and because of uniparental inheritance, these genomes are a useful tool for phylogenetic and evolutionary studies of plants[,,,]. It has been demonstrated that the complete chloroplast genomes is efficient, robust, and authentic for uncovering species discrimination and phylogenetic relationship resolution among species[5,10,]. The chloroplast genomes have been frequently employed to determine the phylogenetic connections of numerous plant species, such as in Amaranthus L.[], chrysanthemums[5,], thanks to the reasonable cost of next-generation sequencing. Therefore, the comparative investigations of the entire chloroplast genome allows us to understand more about the evolution of chloroplast genomes and to identify those important polymorphic loci in plant phylogenetic research[,].

So far, *Rubia cordifolia* is well known and becoming more widely used, while *R. cordifolia* is still mostly found in wild, few work has been reported cultivated products. The continuous excavation may cause the harm of the wild resource of wild *R. cordifolia*, leading to the lack of resources. In this study, the chloroplast genome of *R. alata* and *R. ovatifolia* were analyzed to find a substitute for *R. cordifolia*, and hope to contribute to the improvement of the chloroplast genome of Rubiaceae.

2. Materials and Methods

2.1. Sampling and Sequencing

The Taibai Mountain in Taibai County, Baoji City, Shaanxi Province is where the *R. alata* and *R. ovatifolia* samples were taken. At the collection site, the altitude was 1625 meters, and the latitude and longitude were 107.54 E, 34.06 N. The formal identification of R. alata and R. ovatifolia in this study was

undertaken by Prof. Xinjie Yang (Shaanxi Museum of Traditional Chinese Medicine). Voucher specimens were deposited at the Shaanxi Qinling Application Development and Engineering Center of Chinese Herbal Medicine, Shannxi Province under voucher specimens numbers QLZCY-RA-2302 and QLZCY-RO-2301. The voucher specimens were stored in the Application Development Engineering Technology Research Center of Shaanxi Qinling Chinese Herbal Medicine. Total genomic DNA has been collected from fresh leaves using a modified CTAB method[]. Following electrophoresis detection and concentration measurement, it was refrigerated at -20°C. By applying an ultrasound treatment to the genomic DNA sample, the library was created using purification, amplification through PCR, end repair, A tail, and sequence adapter. The original sequence data was subsequently gathered through the use of Illumina HiSeq sequencing using an Illumina high-throughput sequencing platform.

2.2. Chloroplast Genome Assembly and Annotations

Trimmomatic v0.36[] is employed by Illumina data for eliminating splices and poor-quality data. The subsequent parameter settings are employed: preamble : 20 ; trailing : 20 ; sliding window : 4 : 15 ; minimal lens : 36 ; average qualified : 20. GetCellelle was used for assembling the chloroplast genome after the cleaning of the high-quality Pure Data. The optimised sequence was combined using several Kmer parameters using the splicing software, with lengths set to 107, 117, and 127. The internet-based instrument was used to annotate the completed chloroplast genome CHLOROBOX [](MPI-MP CHLOROBOX - GeSeq (mpg.de). The chloroplast genome map was drawn using OGDRAW version 1.3.1[]. The completed annotated chloroplast genome was uploaded to the NCBI, with *R. alata* (NC_083086.1) and *R. ovatifolia* (NC_083087.1) as the GenBank accession numbers.

2.3. Chloroplast genome features analysis

REPuter v1 was used to determine the relative synonymous codon usage (RSCU) of protein-coding genes[]. The MIcroSAtellite identification tool (Misa) was used to detect simple sequence repeats (SSRs) within the chloroplast genomes of *R. alata* and *R. ovatifolia* species[]. with a threshold of ten for mononucleotide; six for dinucleotide; and other nucleotide repeats are five[].

2.4. Phylogenetic Analysis

The complete chloroplast genome sequences of 102 species belonging to Rubioideae were selected for phylogenetic analysis. MAFFT version 7 was used for multiple sequence alignment[]. MEGA11 software[] was used to construct the neighbor-join (NJ) tree with maximum likelihood and 1000 bootstrap replicates[], *Gentiana scabra* (accession number NC 053842. 1) and *Mostuea* sp. Bremer et al. 5077 (accession number KY378706.1) were selected as outgroups. Chiplot was used to refine the evolutionary tree[].

3. Results

3.1 Characteristics of chloroplast Genomes

In this study, the chloroplast genome positions of *R. alata* of 1–84,802 bp, 84,803 – 111,318 bp, 111,319 – 128,457 bp, and 128,458 – 151,654 bp were large single-copy (LSC) regions, inverted repeat B (IRB) regions, small single-copy (SSC) regions, and inverted repeat A (IRA) regions, respectively. The content of LSC, IRs and SSC in the chloroplast genome was 54.72%, 34.23% and 11.05%, respectively. The chloroplast genome positions of *R. ovatifolia* from 1–84,717 bp, 84,718 – 111,233 bp, 111,234 – 128,350 bp, and 128,351 – 154,866 bp were LSC regions, IRB regions, SSC regions, and IRA regions, respectively. The total G + C content of the whole chloroplast, LSC, SSC, and IR regions of *R. alata* were 37.03%, 34.57%, 30.96%, 42.91%. The total G + C contents of the whole chloroplast, LSC, SSC, and IR regions of *R. ovatifolia* were 37.04%, 34.59%, 30.96%, and 42.92%, respectively (Fig. 1-A, Fig. 1-B, Table 1). In further analysis, the chloroplast genome of *R. alata* contains 85 protein-coding genes, 37 transfer RNA genes (tRNA) and 8 ribosomal RNA genes (rRNA). In addition, the chloroplast genome of *R. ovatifolia* contains 85 protein-coding genes, 37 tRNA and 8 rRNA.(Table 2).

Table 1
Chloroplast genome general features of *R. alata* and *R. ovatifolia*

Characteristics		***R. alata***	***R. ovatifolia***
Size(base pair,bp)		154973	154866
LSC length (bp)		84801	84716
SSC length (bp)		17138	17116
IR length (bp)		26517	26517
Number of genes		130	130
Number of protein-coding genes		85	85
Number of tRNA genes		37	37
Number of rRNA genes		8	8
Intergenic Regions		44689	44627
G + C content	Total (%)	37.03	37.04
	LSC (%)	34.57	34.59
	SSC (%)	30.96	30.96
	IR (%)	42.91	42.92
	CDS (%)	37.58	37.57
	tRNA (%)	55.36	55.36
	rRNA (%)	53.02	53.02
	ALL genes %	39.15	39.16
Protein-coding part (CDS) (% bp)		51.87	51.91
All genes (% bp)		72.22	72.24
Non-coding region (% bp)		27.78	27.76

Table 2
Genes annotated in the chloroplast genome of *R. alata* and *R. ovatifolia*

Category of Genes	Group of Genes	Gene Name (*R. alata*)	Gene Name (R. *ovatifolia*)
Genes for photosynthesis	Subunits of ATP synthase	atpA,B, E, F**, H, I	atpA, B, E, F**, H, I
	Subunits of photosystem II	psbA,B,C, D, E, F, I, J, K, L, M, N, T, Z, ycf3**	psbA, B, C, D, E, F, I, J, K, L, M, N, T, Z, ycf3**
	Subunits of NADH-dehydrogenase	ndhA**, B**,a, C, D, E, F, G, H, I, J, K	ndhA**, B**,a, C, D, E, F, G, H, I, J, K
	Subunits of cytochrome b/f complex	petA, B**, D, G, L, N	petA, B**, D, G, L, N
	Subunits of photosystem I	psaA, B, C, I, J	psaA, B, C, I, J
	Subunit of rubisco	rbcL	rbcL
Self replication	Large subunit of ribosome	rpl14, 16**, 2**,a, 20, 22, 23a, 32, 36	rpl14, 16**, 2**,a, 20, 22, 23a, 32, 36
	DNA dependent RNA polymerase	rpoA, B, C1**, C2	rpoA, B, C1**, C2
	Small subunit of ribosome	rps11,12a, 14, 15, 16**, 18, 19, 2, 3, 4, 7a, 8	rps11, 12a, 14, 15, 16**, 18, 19, 2, 3, 4, 7a, 8
	rRNA genes	rrn16a, rrn23a, rrn4.5a, rrn5a	rrn16a, rrn23a, rrn4.5a, rrn5a
	tRNA genes	trnA-UGC**,a,trnC-GCA,trnD-GUC,trnE-UUC,trnF-GAA,trnG-GCC,trnH-GUG,trnI-CAUa,trnI-GAU**,a,trnK-UUU**,trnL-UAA**,trnL-CAAa,trnL-UAG,trnM-CAUa,trnN-GUUa,trnP-UGG,trnQ-UUG,trnR-UCU,trnR-ACGa,trnS-GCU,trnS-CGA**,trnS-UGA,trnS-GGA,trnT-GGU,trnT-UGU,trnV-UAC**,trnV-GACa,trnW-CCA,trnY-GUA	trnA-UGC**,a,trnC-GCA,trnD-GUC,trnE-UUC,trnF-GAA,trnG-GCC,trnH-GUG,trnI-CAUa,trnI-GAU**,a,trnK-UUU**,trnL-UAA**,trnL-CAAa,trnL-UAG,trnM-CAUa,trnN-GUUa,trnP-UGG,trnQ-UUG,trnR-UCU,trnR-ACGa,trnS-GCU,trnS-CGA**,trnS-UGA,trnS-GGA,trnT-GGU,trnT-UGU,trnV-UAC**,trnV-GACa,trnW-CCA,trnY-GUA

Category of Genes	Group of Genes	Gene Name (*R. alata*)	Gene Name (R. *ovatifolia*)
Other genes	Subunit of Acetyl-CoA-carboxylase	accD	accD
	c-type cytochrom synthesis gene	ccsA	ccsA
	Envelop membrane protein	cemA	cemA
	Protease	clpP**	clpP**
	Maturase	matK	matK
Unkown	Conserved open reading frames	ycf1[a], 2[a], 4	ycf1[a], 2[a],4

** Genes containing introns. [a] Duplicated gene (genes present in the inverted repeat (IR) regions).

3.2 Bias of Codon Usag

Genes from different species or within the same species show different codon usage biases, and the current oscillating use of unbalanced codons in biology is called codon usage bias, helping to better understand environmental adaptability and molecular evolution[]. This study identified 50,299 codons in *all R. alata* protein coding sequences (CDS). The number of Asp amino acids was the highest (1650), and the number of Arg amino acids was the lowest (202). There were 30 codons with RSCU > 1 (Table 3, Fig. 2). In all *R. ovatifolia* CDS, 50,299 codons were identified. The number of amino acids with Tyr was the highest (1479) and the number of amino acids with Ser was the lowest (247). There were 30 codons with RSCU > 1 (Table 4, Fig. 2).

Table 3
codon usage in *R. alata*

Amino acid	Codon	No	RSCU	The codon frequency per amino acid(%)	Amino acid	Codon	No	RSCU	The codon frequency per amino acid(%)
Ala	GCA	710	1.12	28.07	Pro	CCA	563	1.15	28.8
Ala	GCC	385	0.61	15.22	Pro	CCC	398	0.81	20.36
Ala	GCG	290	0.46	11.47	Pro	CCG	284	0.58	14.53
Ala	GCU	1144	1.81	45.23	Pro	CCU	710	1.45	36.32
Cys	UGC	191	0.64	31.77	Gln	CAA	1402	1.52	75.91
Cys	UGU	410	1.37	68.21	Gln	CAG	445	0.48	24.09
Asp	GAC	381	0.38	18.76	Arg	AGA	833	1.69	28.1
Asp	GAU	1650	1.62	81.24	Arg	AGG	368	0.75	12.41
Glu	GAA	2005	1.51	75.68	Arg	CGA	674	1.36	22.74
Glu	GAG	644	0.49	24.31	Arg	CGC	202	0.41	6.81
Phe	UUC	946	0.63	31.5	Arg	CGG	249	0.5	8.4
Phe	UUU	2057	1.37	68.5	Arg	CGU	638	1.29	21.52
Gly	GGA	1191	1.48	36.89	Ser	AGC	247	0.38	6.28
Gly	GGC	353	0.44	10.93	Ser	AGU	731	1.12	18.59
Gly	GGG	595	0.74	18.43	Ser	UCA	733	1.12	18.64
Gly	GGU	1089	1.35	33.73	Ser	UCC	647	0.99	16.46
His	CAC	298	0.49	24.71	Ser	UCG	437	0.67	11.11
His	CAU	908	1.51	75.28	Ser	UCU	1137	1.73	28.92
Ile	AUA	1280	0.92	30.76	Thr	ACA	809	1.27	31.76
Ile	AUC	774	0.56	18.6	Thr	ACC	432	0.68	16.96
Ile	AUU	2107	1.52	50.63	Thr	ACG	310	0.49	12.17
Lys	AAA	2108	1.45	72.72	Thr	ACU	996	1.56	39.1
Lys	AAG	791	0.55	27.29	Val	GUA	951	1.46	36.39
Leu	CUA	764	0.85	14.12	Val	GUC	350	0.54	13.39

Amino acid	Codon	No	RSCU	The codon frequency per amino acid(%)	Amino acid	Codon	No	RSCU	The codon frequency per amino acid(%)
Leu	CUC	365	0.4	6.74	Val	GUG	341	0.52	13.05
Leu	CUG	369	0.41	6.82	Val	GUU	971	1.49	37.16
Leu	CUU	1150	1.28	21.25	Trp	UGG	932	1	99.99
Leu	UUA	1661	1.84	30.69	Tyr	UAC	357	0.39	19.49
Leu	UUG	1103	1.22	20.38	Tyr	UAU	1475	1.61	80.52
Met	AUG	1102	1	100	Stop*	UAA	143	1.26	42.06
Asn	AAC	590	0.47	23.45	Stop*	UAG	106	0.94	31.17
Asn	AAU	1926	1.53	76.55	Stop*	UGA	91	0.8	26.76

Table 4
codon usage in *R. ovatifolia*

Amino acid	Codon	No	RSCU	The codon frequency per amino acid(%)	Amino acid	Codon	No	RSCU	The codon frequency per amino acid(%)
Ala	GCA	708	1.12	28	Pro	CCA	563	1.15	28.8
Ala	GCC	381	0.6	15.07	Pro	CCC	398	0.81	20.36
Ala	GCG	294	0.46	11.62	Pro	CCG	282	0.58	14.42
Ala	GCU	1146	1.81	45.31	Pro	CCU	712	1.46	36.42
Cys	UGC	191	0.64	31.83	Gln	CAA	1408	1.52	76.07
Cys	UGU	409	1.36	68.16	Gln	CAG	443	0.48	23.93
Asp	GAC	379	0.37	18.7	Arg	AGA	839	1.69	28.22
Asp	GAU	1648	1.63	81.3	Arg	AGG	368	0.74	12.38
Glu	GAA	2005	1.52	75.77	Arg	CGA	674	1.36	22.67
Glu	GAG	641	0.48	24.22	Arg	CGC	203	0.41	6.83
Phe	UUC	952	0.63	31.64	Arg	CGG	249	0.5	8.37
Phe	UUU	2057	1.37	68.36	Arg	CGU	640	1.29	21.53
Gly	GGA	1194	1.48	37.04	Ser	AGC	247	0.38	6.27
Gly	GGC	353	0.44	10.95	Ser	AGU	733	1.12	18.62
Gly	GGG	591	0.73	18.34	Ser	UCA	732	1.12	18.59
Gly	GGU	1085	1.35	33.66	Ser	UCC	655	1	16.64
His	CAC	295	0.49	24.56	Ser	UCG	434	0.66	11.02
His	CAU	906	1.51	75.43	Ser	UCU	1136	1.73	28.86
Ile	AUA	1278	0.92	30.74	Thr	ACA	807	1.27	31.71
Ile	AUC	770	0.56	18.52	Thr	ACC	434	0.68	17.05
Ile	AUU	2109	1.52	50.73	Thr	ACG	308	0.48	12.1
Lys	AAA	2100	1.45	72.5	Thr	ACU	996	1.57	39.13
Lys	AAG	797	0.55	27.51	Val	GUA	944	1.44	36.13
Leu	CUA	768	0.85	14.2	Val	GUC	356	0.54	13.62

Amino acid	Codon	No	RSCU	The codon frequency per amino acid(%)	Amino acid	Codon	No	RSCU	The codon frequency per amino acid(%)
Leu	CUC	365	0.4	6.75	Val	GUG	341	0.52	13.05
Leu	CUG	367	0.41	6.78	Val	GUU	972	1.49	37.2
Leu	CUU	1145	1.27	21.17	Trp	UGG	932	1	99.99
Leu	UUA	1657	1.84	30.63	Tyr	UAC	357	0.39	19.45
Leu	UUG	1107	1.23	20.46	Tyr	UAU	1479	1.61	80.56
Met	AUG	1102	1	100	Stop*	UAA	143	1.26	42.18
Asn	AAC	586	0.47	23.27	Stop*	UAG	105	0.93	30.98
Asn	AAU	1932	1.53	76.73	Stop*	UGA	91	0.8	26.84

3.3 Microsatellite Polymorphisms

Simple sequence repeats (SSRs) in the chloroplast genome can be extremely variable at the intra-specific level and are often used as genetic markers in population genetics and evolutionary studies[,]. In this study, we investigated the SSRs in the chloroplast genome of *R. alata* and *R. ovatifolia* (Fig. 3). The chloroplast genome of *R.calata* contains 61 SSRs and the genome of *R. ovatifolia* contains 64 SSRs. The level of mononucleotide repeat content was high (*R. alata*, 78.69%; *R. ovatifolia*, 79.69%) in these two Rubia plant species. These results will provide chloroplast SSR markers that can be used for genetic research, expanding provenance for germplasm selection.

3.4 IR Expansion and Contraction

The IRs serve as integral components of maintaining the stability of the chloroplast genome, as loss of IRs could result in changes in the chloroplast genome[]. In this study, the complete chloroplast genome of Rubiaceae was selected for IR boundary analysis, in order to increase the degree of difference, a sequence of *G. scabra* (accession number:NC053842.1) was added. The results showed that *R. cordifolia*, *R. ovatifolia* and *R. alata* had a very high similarity, the same gene exists at the IR regions, and the expansion and contraction of the gene is not large, which can prove that it has a high genetic relationship. *P. serpens* only had the genes trnR and trnN in the LSC / IRa region, which made it very different from Rubia in the IR regions. At the same time, *Psychotria serpens* was similar to *Coffea racemosa* and *G. scabra* in some cases. *Ophiorrhiza densa* and *Cinchona officinalis* had high variation, especially in *O. densa*, there was gene transformation in JLB, and the same was true in JLA(Fig. 4).

From the above analysis results, it can be seen that the IR boundary region of different Rubiaceae subfamilies has shrunk and expanded during evolution, and there are some differences between species. Generally, the changes in the IR region are small and the chloroplast genome is conservative.

3.5 Comparative analysis of genomic structure

Comparative genome analysis allows the study of how DNA sequences diverge between related species. The complete chloroplast genome sequences of *R. alata* and *R. ovatifolia* were compared with the sequences of other plants of the genus Rubia. The identities of the complete sequences of the seven chloroplast genomes were plotted using the mVISTA *R. cordifolia* annotation as a reference (Fig. 5). The variation of the LSC and SSC regions was significantly greater than that of the IR regions. In summary, the coding regions were more conserved than the non-coding regions. The most divergent coding regions of the seven chloroplast genomes were rps16, psbI-trns-CGA, petN, psbN, ycf3, trnS-GGA, trnL-UAA, petG-psaJ, clpP, rpl16-rps8, ndhE-ndhd, and the distribution of plastid rRNAs were the most conserved.

3.6 Genetic Phylogenetic Analysis

We adopted 102 chloroplast genomes from two tribes of Rubioideae (including *R. alata* and *R. ovatifolia* chloroplast genome newly sequenced in this study). To infer the phylogenetic relationships by using the NJ method. Meanwhile, *G. scabra* (accession number NC 053842.1) and *M.* sp. Bremer et al. 5077 (accession number KY378706.1) were selected as out groups. The NJ tree showed that the selected chloroplast genomes were divided into Rubioideae and outgroups, and the selected Rubioideae sequences were divided into three groups, namely Rubioideae, Psychotrieae alliance, and Spermacoceae alliance. *R. alata* and *R. ovatifolia* chloroplast genomes of this study were assigned to the genus Rubia of the Spermacoceae alliance.The chloroplast genome of *R. alata* was clustered with *R. cordifolia* with a support rate of 66.3%, and *R. ovatifolia* was clustered with *Rubia podantha* with a support rate of 100% (Fig. 6).

4. Discussion

In this study, the complete chloroplast sequences of *R. alata* and *R. ovatifolia* were assembled, annotated and analysed. We then analyzed their chloroplast genome features, codon usage bias, microsatellite polymorphisms, IR expansion and contraction, genome structure comparison, and genetic phylogenetics. The chloroplast genome of *R. alata* and *R. ovatifolia* exhibited a classically four-part structure, similar to the structure of other seed plants[]. The complete chloroplast genome of *R. alata* has a total length of 154,973 bp, with a pair of IRs of 26,517 bp that separate an LSC region of 84,801 bp and an SSC region of 17,138 bp. The complete chloroplast genome of *R. ovatifolia* has a total length of 154,866 bp, with a pair of IRs of 26,517 bp that separate an LSC region of 84,716 bp and an SSC region of 17,116 bp. The highest GC content of *R. alata* and *R. ovatifolia* was 42.91% and 42.92% in the IR region, respectively, indicating that high GC content increases the stability of the IR region and maintains its structure during the evolution of the chloroplast genome[]. The genes of the chloroplast genomes of *R. alata* and *R. ovatifolia* are similar and annotated as different functional genes, but both have unknown function ycf1,

ycf2 and ycf4 genes (Table 2). It is expected that continued research in the future can fully predict chloroplast genome gene function and eliminate these uncertainties.

The codon usage bias in the plant chloroplast genome has been widely used to explore the evolutionary feature of many plant species[]. In this study, 50,299 codons were identified in all protein-coding sequences of *R. alata*. Asp and Arg were the highest and lowest amino acids. Thirty codons were identified with an RSCU > 1. and 50,299 codons were identified in all protein-coding sequences of *R. ovatifolia*.The highest and lowest amino acids were Tyr and Ser. Thirty codons were identified with an RSCU > 1(Fig. 3). This result is similar to the results reported in other angiosperms[,], these features of codon usage preference can help to better understand exogenous gene expression and the evolution mechanisms of the chloroplast genome.

The chloroplast SSR markers are excellent tools for phylogenetic research due to several characteristics, including non-recombination, haploidy, uniparental inheritance, and the low substitution rate[]. They are especially valuable for intraspecific population genetic variation research and interspecific evolutionary and identification studies[,]. A previous study reported that 64 SSRs was identified in the chloroplast genome of *R. podantha*[], The chloroplast genome of *R. alata* and *R. ovatifolia* contained 61 and 64 SSRs, in contrast, in our study. Mononucleotide repeat content was high (*R. alata*, 78.69%; *R. ovatifolia*, 79.69%) in these two Rubia species (Fig. 3). In this study, mononucleotide SSRs were the most abundant and mononucleotide A + T repeat units occupied the highest proportion;these results are consistent with a previous study and verify the hypothesis that chloroplast SSRs are generally composed of short polyadenine (polyA) or polythymine (polyT) repeats and rarely contain tandem guanine (G) or cytosine (C) repeats. Therefore, the chloroplast SSR markers of Rubia can be used to detect the genetic structure, diversity, differentiation and maternity of Rubia, and provide a new way to develop strategies for species conservation, preservation, development of new drug sources and resource development and utilisation.

The size variation in angiosperm plastid genomes is often accompanied by the expansion and contraction of the IR and SSC boundary regions[]. To further resolve the structural evolutionary history of the Rubiaceae chloroplast genomes, we compared the IR/SSC and IR/LSC junctions in eight selected Rubiaceae species, including *R. cordifolia*, *R. alata*, *R. ovatifolia*, *P. serpens*, *O. densa*, *C. officinalis*, *C. racemosa*, *G. scabra*. The chloroplast genomes of *R. alata* and *R. ovatifolia* showed similar JLB, JSB and JSA junction sites. We observed variation in IR/SSC and IR/LSC junctions across Rubiaceae : for example, rps19 gene transferred from LSC to IRb was 14 bp at LSC/IRb junction in *R. cordifolia*, *R. alata* and *R. ovatifolia*, *P.serpens* was 36 bp, *C. racemosa* was 92 bp. *G. scabra* rps19 gene transferred from LSCs to IRb was only 199 bp which was used to increase difference, *O. densa* and *C. officinalis* gene transfers occurred from LSCs to IRb.(Fig. 4). The results showed that *R. cordifolia*, *R. ovatifolia* and *R. alata* have a very high similarity, the same gene exists at the IR regions, and the expansion and contraction of the gene is not large, which can prove that it had a high genetic relationship. Meanwhile, the results also showed that IR expansion and contraction can also provide a reference for plant origin and phylogenetic relationship.

With at least ten different variable regions found in the alignment of the 7 chloroplast genome sequences, based on mVISTA in the chloroplast genome of Rubia genus (Fig. 5), furthermore, *R. cordifolia* as a traditional medicine for cooling blood, plays an important role in traditional Chinese medicine[4]. novel indel and repeat markers could be developed to aid in species identification and authentication of Rubia species.

This study is based on previous studies on the chloroplast genome of *R. podantha*[49], the complete sequence of 104 dicotyledonous chloroplast genomes was downloaded from NCBI. The phylogenetic analysis based on the complete chloroplast sequences showed that the selected Rubioideae sequences were divided into three groups, namely Rubioideae, Psychotrieae alliance and Spermacoceae alliance. The chloroplast genomes of *R. alata* and *R. ovatifolia* were assigned to the genus Rubia of the Spermacoceae alliance, the chloroplast genome of R.alata clustered with *R. cordifolia* with 66.3% support, and *R. ovatifolia* clustered with *R. podantha* with 100% support (Fig. 6). The results showed that *R. alata* and *R. ovatifolia* and *R.cordifolia* could be classified into a single branch with a high degree of support. Our results will provide more information on Rubia for molecular assisted breeding.

In China, *R. cordifolia* has a high medical and commercial value and there is also a great deal of research on it. However, *R. alata* and *R. ovatifolia*, which belong to the same genus of Rubiaceae, have been neglected so far. Based on the analysis results of this study, we hope to increase the attention of *R. alata* and *R. ovatifolia*, which are two endemic plants in China, and at the same time, add the study of chloroplast genome of some Rubiaceae plants in order to complete the research of Rubiaceae at the chloroplast genome level, and provide theoretical basis and reference for germplasm identification, new drug development and resource utilization..

Conclusions

This study completed the annotation and general analysis of the chloroplast genome of *R. alata* and *R. ovatifolia*, and added the research on the chloroplast genome of Rubiaceae plants, in order to provide help for the identification and valorisation of Rubiaceae plants.

Abbreviations

LSC	Large single copy
SSC	Small single copy
IRs	Inverted repeats
IRA	Inverted repeat A
IRB	Inverted repeat B
bp	base pairs

tRNA transfer RNA

rRNA Ribosomal RNA

CDS Protein coding sequences

NJ Neighbor-Join

Declarations

Authors’contributions

LP and JZS conceived the study and funded it. JZS conducted data analysis and drafted an earlier version of the manuscript. XYC, YYJ, YGY, GZ and YBY are involved in project management and assistance, and all authors endorse the final manuscript.

Funding

This work was funded by Natural Science Foundation of Shaanxi Province, grant number 2021-JQ-733; Shaanxi Provincial Administration of Traditional Chinese Medicine Project, grant number 2021-QYZL-02.

Data availability

The data supporting the findings of this study are freely available in GenBank on the NCBI website at National Center for Biotechnology Information (nih.gov),using the accession number NC_083086.1 and NC_083087.1.

Ethics approval and consent to participate

The authors confirm that all methods comply with local and national regulations.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Acknowledgements

Thanks to all the teachers and students who had a hand in the design of this work. We finally thank the anonymous reviewers for reviewing the manuscript and providing valuable comments.

References

1. Chen WQ, Luo XR, Gao YZ, Ruan YZ: Flora of China: Rubiaceae. Beijing. 1999, 71 (1): 287–318.
2. Xu K, Wang P, Wang L, Liu C, Xu S, Cheng Y, Wang Y, Li Q, Lei H. Quinone derivatives from the genus Rubia and their bioactivities. Chem Biodivers. 2014 Mar;11(3):341–63.
3. Zhao SM, Kuang B, Fan JT, Yan H, Xu WY, Tan NH. Antitumor cyclic hexapeptides from rubia plants: history, chemistry, and mechanism (2005–2011). Chimia (Aarau). 2011;65(12):952–6.
4. Chinese Pharmacopoeia Commission: Chinese Pharmacopoeia. Part I. 2020, Beijing, 245.
5. Chen WQ, Luo XR, Gao YZ, Ruan YZ: Flora of China: Rubiaceae. Beijing. 1999, 71 (2): 311.
6. Zhao SM, Wang Z, Zeng GZ, Song WW, Chen XQ, Li XN, Tan NH. New cytotoxic naphthohydroquinone dimers from Rubia alata. Org Lett. 2014 Nov 7;16(21):5576–9.
7. Chen WQ, Luo XR, Gao YZ, Ruan YZ: Flora of China: Rubiaceae. Beijing.1999, 71 (2):306.
8. Jensen PE, Leister D. Chloroplast evolution, structure and functions. F1000Prime Rep. 2014 Jun 2;6:40.
9. Tsudzuki T, Wakasugi T, Sugiura M. Comparative analysis of RNA editing sites in higher plant chloroplasts. J Mol Evol. 2001 Oct-Nov;53(4–5):327–32.
10. Palmer, J.D. The Molecular Biology of Plastids; Academic Press: Pittsburgh, PA, USA, 1991; CHAPTER 2-Plastid chromosomes: Structure and evolution.
11. Jansen R.K, Raubeson L.A,Boore J.L, Depamphilis C.W, Chumley T.W, Haberle R.C, Wyman S.K, Alverson A.J, Peery R, Herman S.J, et al. Methods for Obtaining and Analyzing Whole Chloroplast Genome Sequences. Methods Enzymol. 2005, 395, 348–384.
12. Ma Y, Zhao L, Zhang W, Zhang Y, Xing X, Duan X, Hu J, Harri A.J, Liu P, Dai S, et al. Origins of cultivars of Chrysanthemum-Evidence from the chloroplast genome and nuclear LFY gene. J. Syst. Evol. 2020, 58, 925–944.
13. Li Y, Zhang J, Li L, Gao L, Xu J, Yang M. Structural and Comparative Analysis of the Complete Chloroplast Genome of Pyrus hopeiensis-"Wild Plants with a Tiny Population"-and Three Other Pyrus Species. Int J Mol Sci. 2018 Oct 20;19(10):3262.
14. Ahmed I, Biggs P.J, Matthews P.J, Collins L.J, Hendy M.D, Lockhart P.J. Mutational dynamics of aroid chloroplast genomes. Genome Biol. Evol. 2012, 4, 1316–1323.
15. Song W, Chen Z, He L, Feng Q, Zhang H, Du G, Shi C, Wang S. Comparative chloroplast genome analysis of wax gourd (Benincasa hispida) with three Benincaseae Species, revealing evolutionary dynamic patterns and phylogenetic implications. Genes 2022, 13, 461.
16. Jansen R.K, Cai Z, Raubeson L.A, Daniell H, Depamphilis C.W, Leebens-Mack J, Müller K.F, Guisinger-Bellian M, Haberle R.C, Hansen A.K, et al. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc. Natl. Acad. Sci. USA 2007, 104, 19369–19374.
17. Zhang N, Erickson D.L, Ramachandran P, Ottesen A.R, Timme R.E, Funk V.A, Luo Y, Handy S.M. An analysis of Echinacea chloroplast genomes: Implications for future botanical identification. Sci. Rep. 2017, 7, 216.

18. Zhang X, Zhang H, Landis J, Deng T, Meng A,Sun H, Peng Y, Wang H, Sun Y. Plastome phylogenomic analysis of Torreya (Taxaceae). J. Syst. Evol. 2019, 57, 607–615.
19. Wang H, Moore M, Barrett R, Landrein S, Sakaguchi S, Maki M, Weng H, Wang H. Plastome phylogenomic insights into the Sino–Japanese biogeography of Diabelia (Caprifoliaceae). J. Syst. Evol. 2020, 58, 972–987.
20. Nguyen VB, Linh Giang VN, Waminal NE, Park HS, Kim NH, Jang W, Lee J, Yang TJ. Comprehensive comparative analysis of chloroplast genomes from seven Panax species and development of an authentication system based on species-unique single nucleotide polymorphism markers. J Ginseng Res. 2020 Jan;44(1):135–144.
21. Viljoen E, Odeny DA, Coetzee MPA, Berger DK, Rees DJG. Application of Chloroplast Phylogenomics to Resolve Species Relationships Within the Plant Genus Amaranthus. J Mol Evol. 2018 Apr;86(3–4):216–239.
22. Xu Y, Liao B, Ostevik K.L, Zhou H, Wang F, Wang B, Xia H. The Maternal donor of Chrysanthemum cultivars revealed by comparative analysis of the chloroplast genome. Front. Plant Sci. 2022, 13, 923442.
23. Mehmood F, Rahim A, Heidari P, Ahmed I, Poczai P. Comparative plastome analysis of Blumea, with implications for genome evolution and phylogeny of Asteroideae. Ecol. Evol. 2021, 11, 7810–7826.
24. Henriquez C.L, Ahmed I, Carlsen M.M, Zuluaga A, Croat T.B, McKain M.R. Molecular evolution of chloroplast genomes in Monsteroideae (Araceae). Planta 2020, 251, 72.
25. Mchloroplastherson H, van der Merwe M, Delaney SK, Edwards MA, Henry RJ, McIntosh E, Rymer PD, Milner ML, Siow J, Rossetto M. Capturing chloroplast variation for molecular ecology studies: a simple next generation sequencing approach applied to a rainforest tree. BMC Ecol. 2013 Mar 14;13:8.
26. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114–20.
27. Zhang D, Gao F, Jakovlić I, Zou H, Zhang J, Li WX, Wang GT. PhyloSuite: An integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. Mol Ecol Resour. 2020 Jan;20(1):348–355.
28. Greiner S, Lehwark P, Bock R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Res. 2019 Jul 2;47(W1):W59-W64.
29. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 2001 Nov 15;29(22):4633–42.
30. Beier S, Thiel T, Münch T, Scholz U, Mascher M. MISA-web: a web server for microsatellite prediction. Bioinformatics. 2017 Aug 15;33(16):2583–2585.
31. Thiel T, Michalek W, Varshney RK, Graner A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). Theor Appl Genet. 2003 Feb;106(3):411–22.

32. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform. 2019 Jul 19;20(4):1160–1166.
33. Tamura K, Stecher G, Kumar S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Mol Biol Evol. 2021 Jun 25;38(7):3022–3027.
34. Prillo S, Deng Y, Boyeau P, Li X, Chen PY, Song YS. CherryML: scalable maximum likelihood estimation of phylogenetic models. Nat Methods. 2023 Aug;20(8):1232–1236.
35. Xie J, Chen Y, Cai G, Cai R, Hu Z, Wang H. Tree Visualization By One Table (tvBOT): a web application for visualizing, modifying and annotating phylogenetic trees. Nucleic Acids Res. 2023 Jul 5;51(W1):W587-W592.
36. Qian R, Ye Y, Hu Q, Ma X, Zheng J. Complete Chloroplast Genome of Gladiolus gandavensis (Gladiolus) and Genetic Evolutionary Analysis. Genes (Basel). 2022 Sep 7;13(9):1599.
37. Mohammad-Panah N, Shabanian N, Khadivi A, Rahmani M.-S, Emami A. Genetic structure of gall oak (Quercus infectoria) characterized by nuclear and chloroplast SSR markers. Tree Genet. Genomes 2017, 13, 70–82.
38. Park S, Sang I.P, Gil J, Um Y, Jung C.S, Lee J.H, Kim S.C, Kim H.B, Lee Y, Development of Simple Sequence Repeat SSR Markers Based on Chloroplast DNA to Distinguish 3 Angelica Species. Korean Soc. Hortic. Sci. 2016, 10, 226.
39. Lee E, Smith D, Fanwick P, Byrn, S. Characterization and Anisotropic Lattice Expansion/Contraction of Polymorphs of Tenofovir Disoproxil Fumarate. Cryst. Growth. Des. 2010, 10, 2314–2322.
40. Shen X, Guo S, Yin Y, Zhang J, Yin X, Liang C, Wang Z, Huang B, Liu Y, Xiao S, Zhu G. Complete Chloroplast Genome Sequence and Phylogenetic Analysis of Aster tataricus. Molecules. 2018 Sep 21;23(10):2426.
41. Zhao Y, Qu D, Ma Y. Characterization of the Chloroplast Genome of Argyranthemum frutescens and a Comparison with Other Species in Anthemideae. Genes (Basel). 2022 Sep 25;13(10):1720.
42. Qin Z, Cai Z, Xia G, Wang M. Synonymous codon usage bias is correlative to intron number and shows disequilibrium among exons in plants. BMC Genomics. 2013 Jan 28;14:56.
43. Liu HY, Yu Y, Deng YQ, Li J, Huang ZX, Zhou SD. The Chloroplast Genome of Lilium henrici: Genome Structure and Comparative Analysis. Molecules. 2018 May 26;23(6):1276.
44. Jian HY, Zhang YH, Yan HJ, Qiu XQ, Wang QG, Li SB, Zhang SD. The Complete Chloroplast Genome of a Key Ancestor of Modern Roses, Rosa chinensis var. spontanea, and a Comparison with Congeneric Species. Molecules. 2018 Feb 12;23(2):389.
45. Ebert D, Peakall R. Chloroplast simple sequence repeats (chloroplastSSRs): technical resources and recommendations for expanding chloroplastSSR discovery and applications to a wide array of plant species. Mol Ecol Resour. 2009 May;9(3):673–90.
46. Huang LS, Sun YQ, Jin Y, Gao Q, Hu XG, Gao FL, Yang XL, Zhu JJ, El-Kassaby YA, Mao JF. Development of high transferability chloroplastSSR markers for individual identification and genetic investigation in Cupressaceae species. Ecol Evol. 2018 Apr 20;8(10):4967–4977.

47. Pan L, Li Y, Guo R, Wu H, Hu Z, Chen C. Development of 12 chloroplast microsatellite markers in Vigna unguiculata (Fabaceae) and amplification in Phaseolus vulgaris. Appl Plant Sci. 2014 Mar 4;2(3):apps.1300075.
48. Zhao SY, Muchuku JK, Liang HY, Wang QF. A complete chloroplast genome of a traditional Chinese medicine herb, Rubia podantha, and phylogenomics of Rubiaceae. Physiol Mol Biol Plants. 2023 Jun;29(6):843–853.
49. Drescher A, Ruf S, Calsa T Jr, Carrer H, Bock R. The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. Plant J. 2000 Apr;22(2):97–104.

Figures

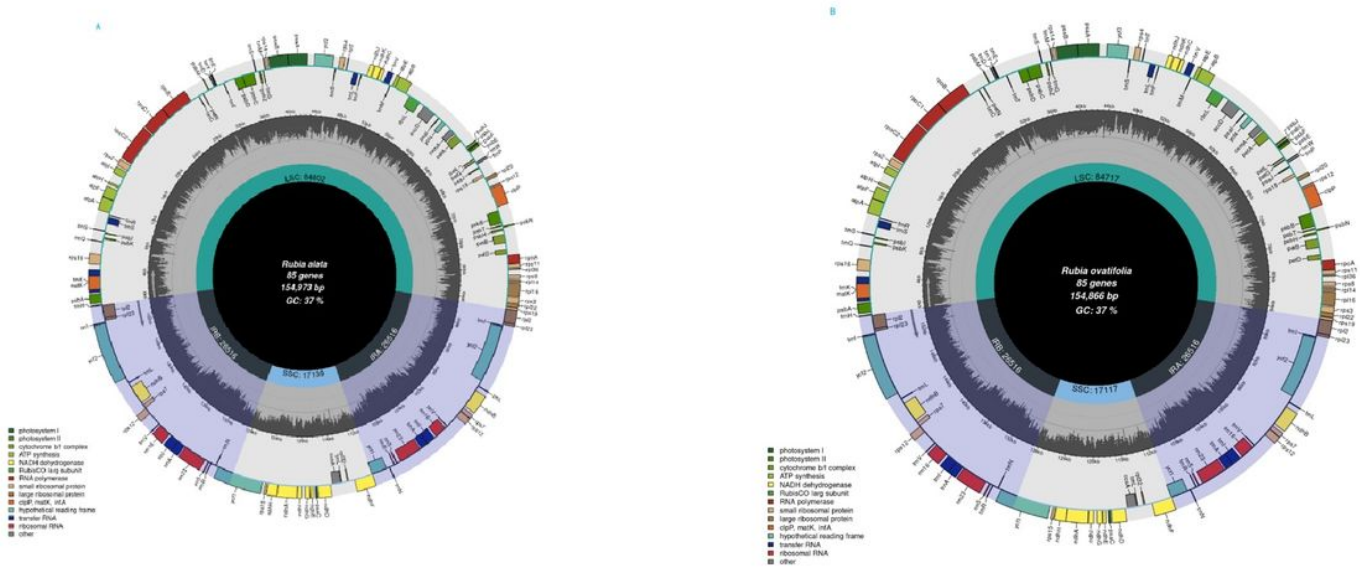

Figure 1

The chloroplast genome of *R. alata* and *R. ovatifolia*. Fig 1-A is the chloroplast genome circle diagram of *R.alata*, Fig 1-B is the chloroplast genome circle diagram of *R. ovatifolia*. Genes drawn outside the circle are shown clockwise, while those inside the circle are shown counterclockwise. Additionally, genes with different functions are coloured..

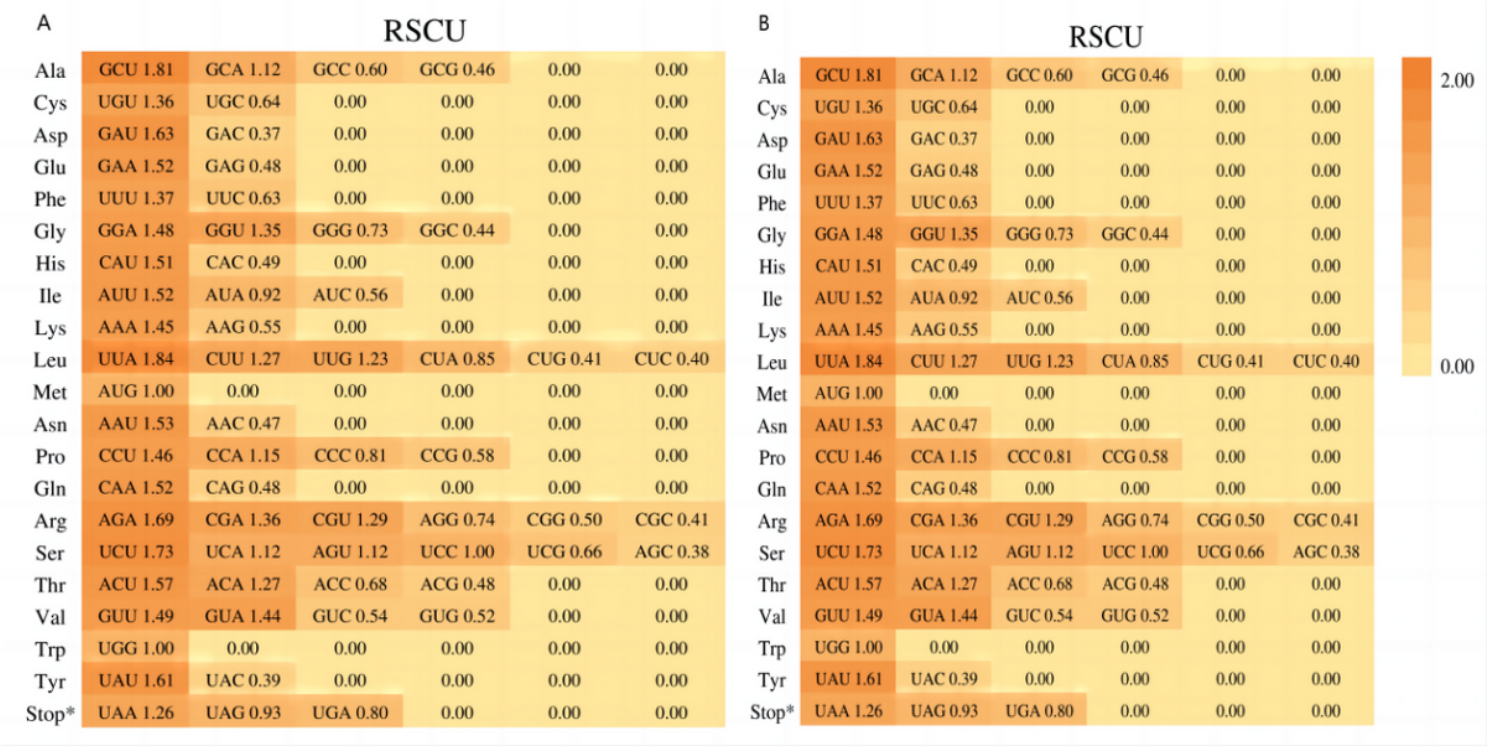

Figure 2

In this study, heat maps were used to represent RSCU values: Figure A shows the RSCU value of the codons in *R. alata*and Figure B shows the RSCU value of the codon in *R. ovatifolia*.

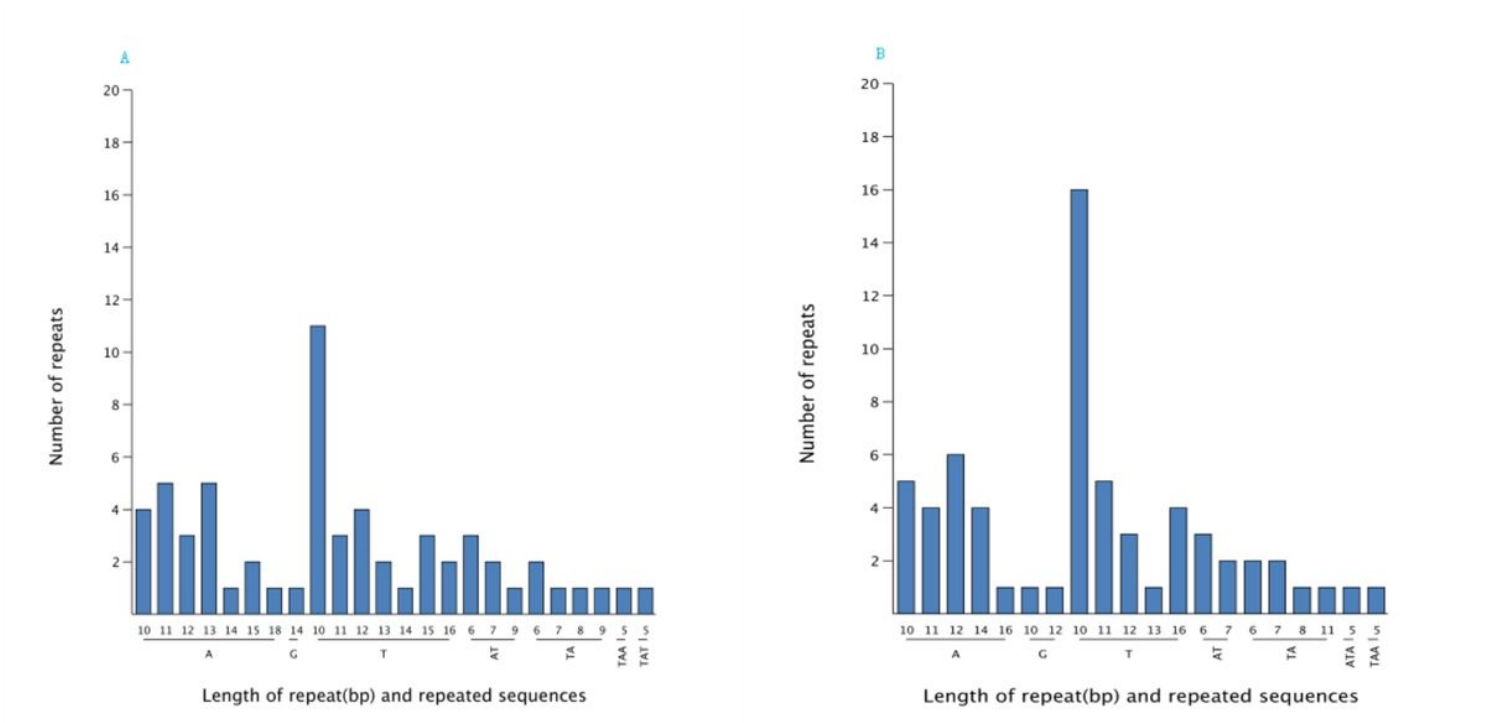

Figure 3

The number of repeated SSRs in the chloroplast genomes of *R. alata* and *R. ovatifolia* are shown in the figure. Fig.A shows the SSRs in *R. alata* and Fig.B shows the SSRs in *R. ovatifolia*.The x-axis represents the type of repetition and the y-axis is the number of repetitions.

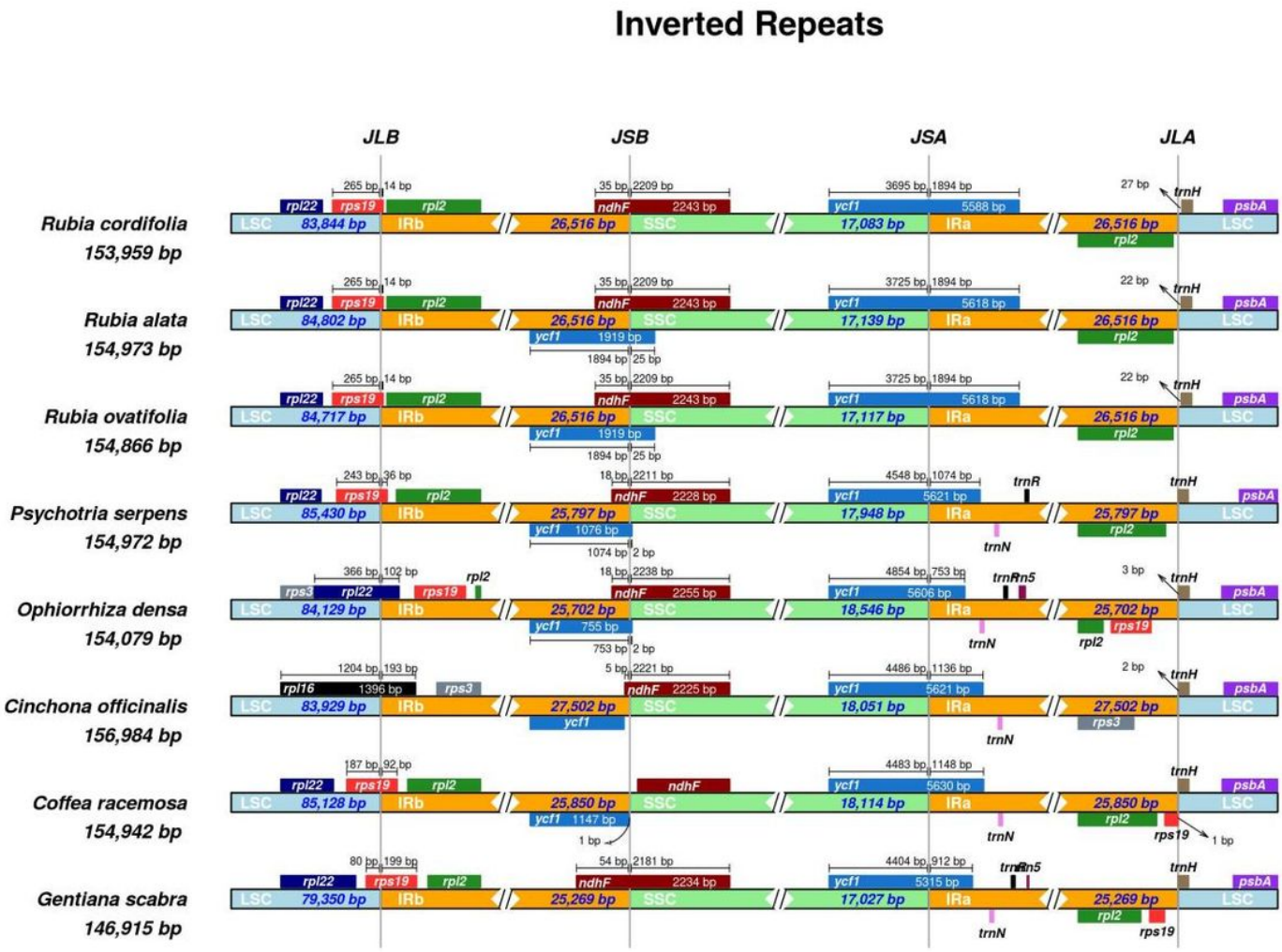

Figure 4

The selected IR boundary analysis sequences are: *R. cordifolia* (OK326894), *R. alata* (NC083086), *R. ovatifolia*(NC083087), *P. serpens* (NC069807), *O. densa* (NC058252), *C. officinalis*(MZ151891), *C. racemosa* (MW970412), *G. scabra* (NC053842). The genes located at the IRa/b junctions are represented by the coloured boxes above the horizontal line (sense), while the gene segments are represented by the boxes below the horizontal line (antisense).

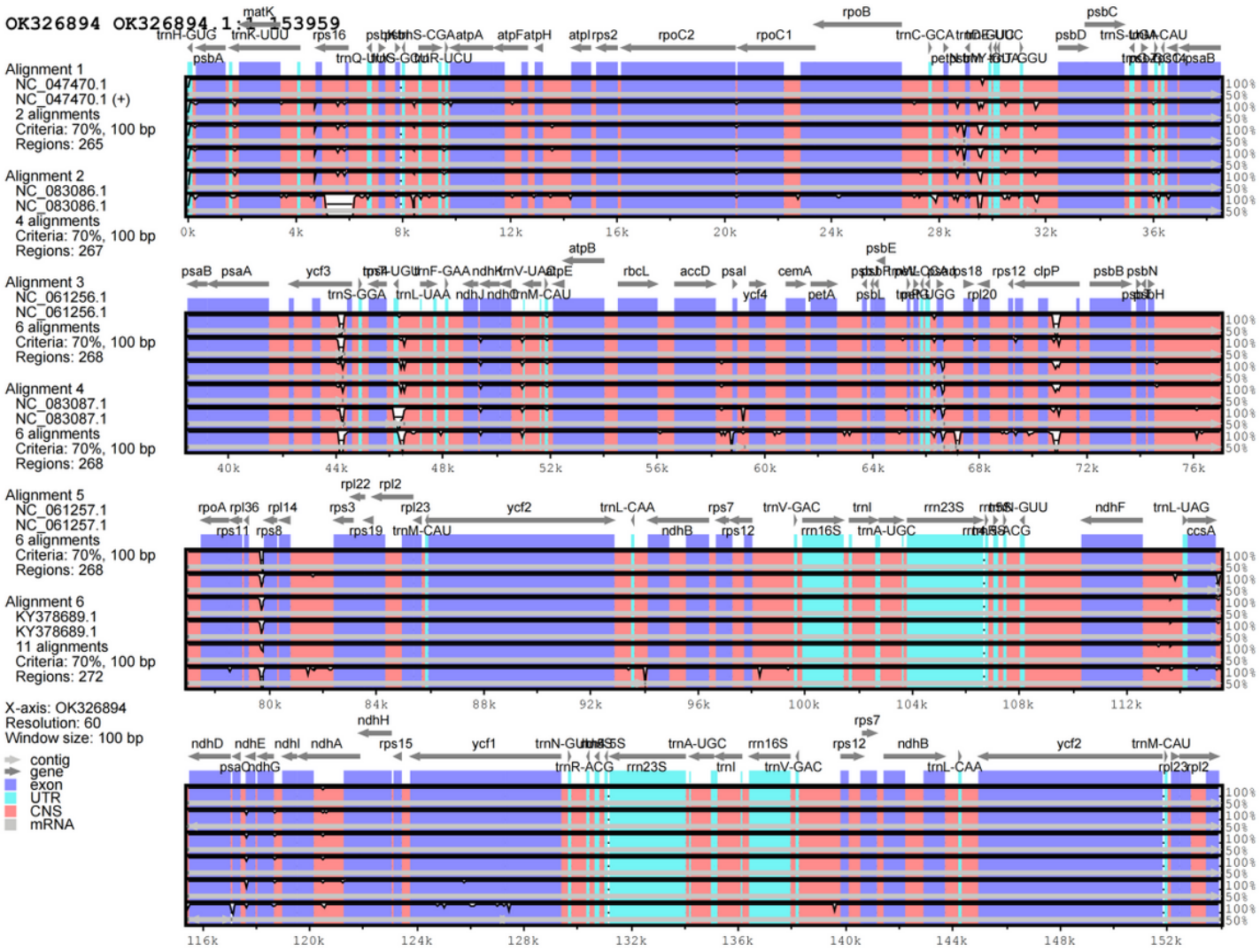

Figure 5

Comparison of the chloroplast genome sequences of *R. cordifolia*, *R. alata*, *R. ovatifolia*, *R. podantha*, *R. yunnanensis* and *R. horrida* generated with mVISTA. Among them, *R. cordifolia*with accession number OK326894 is used for sequence annotation.The grey arrow indicates the position and direction of each gene. The red and blue regions represent the intergenic region and the gene region, respectively. The vertical scale represents a similar percentage, with a range of 50 % -100 %.

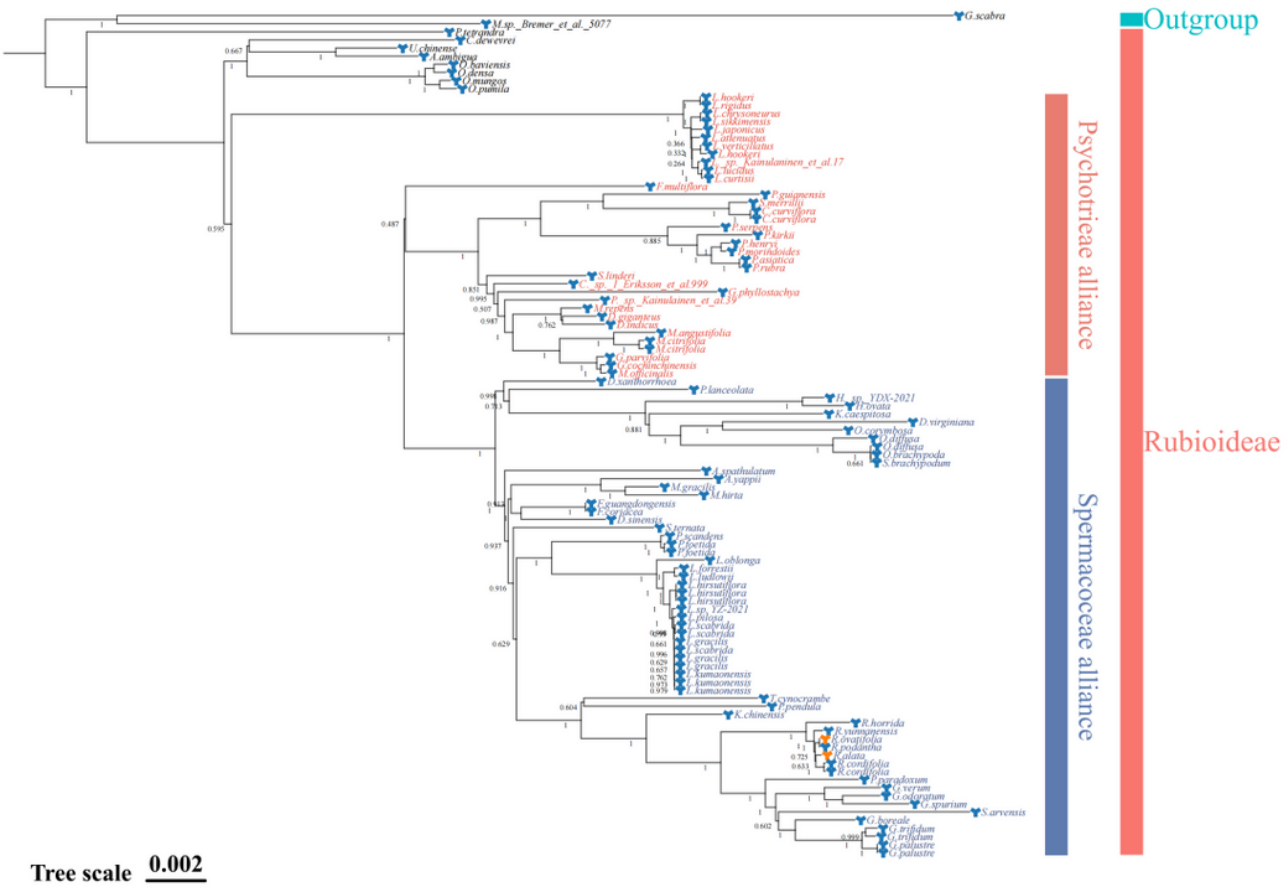

Figure 6

The phylogenetic tree was constructed based on the 104 dicotyledoneae chloroplast genomes: two orange symbolWye which was sequenced in this study. The Maximunm Composite Likelihood model of NJ was used in this study.The bootstrap value was equalled to 1000.