The Complete Chloroplast Genome Sequence of *Clerodendranthus Spicatus*, A Medicinal Plant For Preventing And Treating Kidney Diseases From Lamiaceae Family

Qing Du (1717653300@qq.com)
Qinghai Nationalities University https://orcid.org/0000-0002-0732-3377
Mei Jiang
Chinese Academy of Medical Sciences & Peking Union Medical College Institute of Materia Medica
Sihui Sun
CAMS PUMC IMPLAD: Chinese Academy of Medical Sciences & Peking Union Medical College Institute of Medicinal Plant Development
Liqiang Wang
Heze University
Shengyu Liu
CAMS PUMC IMM: Chinese Academy of Medical Sciences & Peking Union Medical College Institute of Materia Medica
Chuanbei Jiang
Genepioneer Biotechnologies Inc.
Haidong Gao
Genepioneer Biotechnologies Inc
Haimei Chen
CAMS PUMC IMM: Chinese Academy of Medical Sciences & Peking Union Medical College Institute of Materia Medica
Chang Liu
CAMS PUMC IMM: Chinese Academy of Medical Sciences & Peking Union Medical College Institute of Materia Medica

Research Article

Keywords: Clerodendranthus spicatus, Chloroplast genome, Hypervariable region, Comparison of regions and genes, Phylogenetic analysis, Codon usage

Posted Date: June 23rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-497985/v1

Version of Record: A version of this preprint was published at Molecular Biology Reports on January 21st, 2022. See the published version at https://doi.org/10.1007/s11033-022-07135-4.

Abstract

Clerodendranthus spicatus (Thunb.) C.Y.Wu is one of the most important medicine for the treatment of nephrology which distributes in south-east of China. In this study, we obtained the complete chloroplast genome of *C. spicatus* with a length of 152155bp, including a large single copy (LSC) region of 83098bp, small single copy (SSC) region of 17665bp and a pair of inverted repeat (IR) regions of 25696bp with the GC content of 37.86%. The genome contains 36 tRNA, 8 rRNA and 87 protein-coding genes. Most of them have one intron except the *ycf*3, *rps*12 and *clp*P genes. The length of rRNAs varies from 131bp to 2811bp and the GC contents are between 45.28% and 56.54%. The frequency of Isoleucine is fruitful accounting for 4.17%. The codons of AUG, UUA and AGA codon had presence of higher codon usage bias. For the repetitive sequence analysis, Thirty-six tandem repeats were identified with certain conditions. Forty interspersed repeats were identified, including 22 palindromic repeats and 18 direct repeats. The diverse positions of the specific *rps*19, *ycf*1, *rpl*2, *trn*H, *psb*A genes within the IR boundary analysis. The genetic distance analysis of the intergenic spacer regions for 5 relative species showed the areas of *ndh*G-*ndh*I, *acc*D-*psa*I, *rps*15-*ycf*1, *rpl*20-*clp*P, *ccs*A-*ndh*D had high K2p value to distinguish the species through developing the molecular markers. From phylogenetic tree, *C. spicatu* was closely related to the genus of two *Salvia* speices, *Tectona grandis, Cistanche deserticola* and *Glechoma longituba* belonged to the *Lamiales*.

Introduction

Clerodendranthus spicatus (Thunb.) C.Y.Wu, commonly called as Yanumiao, Maoxugong and Maoxucao [1]. It is a kind of perennial herbs and is the first reported among Clerodendranthus genus in the Lamiaceae family. It is mainly produced in the area of Guangdong, Hainan, Guangxi, Yunnan, Fujian province and Taiwan province. And it is also distributed in India, Myanmar, Thailand, Indonesia, Philippines to Australia and adjacent islands [2]. *C.spicatus from the* Wild fields commonly grows in the wet place under the forest and mostly are cultivated up to 1050 meters above sea level [3]. *C. spicatus* has numerous chemical constituents, such as flavonoids [4], phenols [5–6], terpenes [7], volatile oils [8] and lignans [9]. The entire plants or aerial parts are widely used in the treatment of chronic nephritis, cystitis, lithangiuria and rheumatoid arthritis, which hold good effects on the kidney disease [10–11].

The chloroplasts are important organelles in the plant photosynthesis [12]. The study of chloroplast genome is of great significance to reveal the mechanism and metabolic regulation of plant photosynthesis [13]. At the same time, a large number of proteins in chloroplast come from nuclear genome, which is only a semi-autonomous organelle [14]. Further study of chloroplast genome helps to understand the interaction between nuclear genome and chloroplast genome [15]. Moreover, chloroplast genomes have the important value in revealing the plants origin, molecular evolution system and relationship between different species [16]. The chloroplast genome sequences of *C. spicatu* have not been reported so far. In the study, the chloroplast genome of *C. spicatus* were sequenced and annotated for the first time. After its structure features were characterized and analyzed, we chose other 5 species (*Cistanche deserticola, Glechoma longituba, Salvia miltiorrhiza, Salvia miltiorrhiza f. alba and Tectona*

grandis) from Lamiaceae family and 1 species (*Epimedium brevicornu*) from Berberidaceae as the outgroup for the further phylogenetic investigations regarding the divergences and genetic evolution (Table S1) [17]. These seven species have the common medical values related to urinary system and are used to treat the renal diseases, such as acute and chronic nephritis, cystitis, urinary calculi, lithangiuria, nephropyelitis, diuresis, tonifying the kidney and Yang, premature ejaculation, soreness of waist and skelalgia.

The genetic clustering and genetic similarity analysis of 87 *C. spicatus* samples by ISSR marker technique have been reported [18]. The results showed that there were 120 polymorphic bands, and the UPGM A method was used to gather the materials into 2 categories for further proving them. The ISSR markers are feasible for the genetic diversity analysis of *C. spicatus* resources. ISSR markers can well reveal the genetic differences and genetic relationships in *C. spicatus* germplasm resources to reveal their genetic characteristics. It provides theoretical basis for species identification, variety selection, product development and application of *C. spicatus*. We carried out search the certain genes from hypervariable regions for direct cloning and discriminating 5 relative species based on the results of phylogenetic trees [19]. The analysis data of *C. spicatus* chloroplast genome can be meaningful for further identification and researches regarding genetic expression.

Materials And Methods

2.1 Plant materials

Fresh leaves of *C. spicatus* were collected from the Guangxi Medical Botanical Garden, Nanning, Guangxi, China (Geospatial coordinates: N22.859968, E108.383475). The voucher specimen was deposited at the Institute of Medicinal Plant Development with the specimen number is implad 201910126 (Contact person: HM Chen; Email: hmchen@implad.ac.cn).

2.2 DNA extraction and detection of DNA quality

The fresh leaves of *C. spicatus* was collected and treated by Silica gel immediately. Total genomic DNA was extracted from the dried leaves using the plant genomic DNA kit (Tiangen Biotech, Beijing, China). The quality of the extracted genome DNA was detected by 1% agarose gel electrophoresis and microspectrophotometer [20].

2.3 Chloroplast genome sequencing, assembly, and annotation

DNA extracts were fragmented for 300 bp short-insert library construction and sequenced 2 × 150 bp paired-end reads on an Illumina Solexa sequencing platform (Hiseq 2500, San Diego, CA, USA) [21]. The raw reads were filtered by using the Trimmomatic 0.35 to remove adapters and low-quality bases [22]. Then, about 5.0 GB clean reads were assembled by using NOVOPlasty (v4.2) with default parameters (Type = chloroplast, K-mer = 39, Read Length = 150, Insert size = 300, Single/Paired = PE, Insert Range = 1.9) and annotated through the CpGAVAS2 web service (http://www.herbalgenomics.org/cpgavas/) [23–24]. The annotation included the prediction of protein-coding genes, tRNA genes, rRNA genes and repeat

sequence analysis. The annotations with problems of chloroplast genome were manually corrected using the Apollo software [25]. At last, the assembly and annotation results of Clerodendranthus spicatus chloroplast genome was deposited in GenBank with the accession number MZ063774.

2.4 Characteristics analysis of specific genes and proteins

From the circular map of chloroplast genome, schematic presentation and structures of cis-splicing genes, trans-splicing genes and rRNA were visualized and depicted using CPGview-RSG software (http://www.herbalgenomics.org/cpgview/). The genes contents, lengths of introns and exons in genes and features of various repeat sequences were visualized.

2.5 Relationship between codon usage and tRNA types

In chloroplast genome of *Clerodendranthus spicatus*, tRNA types and codon usages were given the results through the annotated software.

2.6 Analysis of relative synonymous codon Usage (RSCU) and Codon Adaptation Index

Based on the data of the chloroplast genomes of *C.* spicatus, the relative synonymous codon usage (RSCU) was calculated as the ratio of the observed frequency of a codon to the expected frequency of the same codon within a synonymous codon group in the entire coding sequence of the gene concerned. The RSCU patterns and quantity was estimated by using CodonW 1.4.4 annotation software (http://codonw.sourceforge.net/) [26]. The Codon Adaptation Index (CAI) value (https://www.bioinformatics.nl/emboss-explorer/) [27] was calculated.

2.7 Repeats studies

Regarding chloroplast genome of *Clerodendranthus spicatus*, the identification of interspersed repeats and tandem repeats can be analyzed using the tool of REPuter (parameter:Hamming distance = 3, Minimum repetition unit = 8bp, https://bibiserv.cebitec.uni-bielefeld.de/reputer/) [28], Tendem Repeat Finder (TRF4.09) (parameter: matching weight = 2, mismatching penalty = 7, Delta = 7, match probability = 80, PI = 10, Minscore = 50, MaxPeriod = 500, http://tandem.bu.edu/trf/trf.html) [29] and MISA (parameter: unit_size,min_repeats = 1–10, 2–6, 3–5, 4–5, 5–5, 6 – 5; max_difference_between_2_SSRs = 100, http://pgrc.ipk-gatersleben.de/misa/) [30] for the identification of simple repeats, respectively.

2.8 Comparison of the LSC, SSC, and IR region borders and existing characteristic genes

The LSC, SSC, and IR regions boundary and sizes of *C. spicatus* chloroplast genome was visualized by the IRscope software (https://irscope.shinyapps.io/irapp/),compared with five closely related from Lamiales (*Cistanche deserticola*, *Glechoma longituba*, *Salvia miltiorrhiza*, *Salvia miltiorrhiza f. alba* and *Tectona grandis*) and one species from different family of Berberidaceae (*Epimedium brevicornu*) [31]. These species selected have the similar effects on the renal function and urinary system of body.

Meanwhile same or differentia genes existing in the different areas and junction sites were comparatively analyzed shown in the graph.

2.9 Hypervariable analysis of related species

The genetic distances of intergenic spacer regions were calculated by using the distmat program from EMBOSS (v6.3.1) [32] with the Kimura 2-parameters (K2p) [33] evolutionary model among these 5 related species from Lamiaceae family including varieties of *Clerodendranthus spicatus*, *Glechoma longituba*, *Salvia_miltiorrhiza*, *Salvia_miltiorrhiza f. alba* and *Tectona grandis.*

2.9 Phylogenetic analysis

The assembled chloroplast genomes of C. spicatus and other 14 reported plastomes, including the 6 species belonged to the *Lamiales* (*Clerodendranthus spicatus, Glechoma longituba, Salvia miltiorrhiza, Salvia miltiorrhiza f. alba,Tectona grandis and Cistanche deserticola*), *Eucommia ulmoides* from *Eucommiaceae* family, *Dipsacus asper* from *Caprifoliaceae* family, *Cullen corylifolium* from *Fabaceae* family, four species from *Polygonaceae* family (*Rheum officinale, Rheum palmatum, Rheum tanguticum* and *Rheum nobile*), *Epimedium brevicornu* from *Berberidaceae* family and *Dioscorea polystachya* from *Dioscoreaceae* family as the outgroup (Table S1), were downloaded from GenBank database (Table S1). The shared sequences coding for amino acids in protein were extracted, concatenated and aligned by the PhyloSuite (v 1.2.2) [34] coupled with the MAFFT (v 7.313) [35]. Phylogenetic analysis was conducted based on maximum likelihood (ML) analyses implemented in IQ-TREE (v1.6.8) [36] under the TVM + F + I + G4 nucleotide substitution model. The significance level for the phylogenetic tree was assessed by bootstrap analysis with 1000 replications. The phylogenetic tree was visualized using the MEGA 5 [37]. In addition, we carried out the common, different and sole proteins among the shared CDS-AA extracted from the 15 species so as to discuss the differences among species.

Results And Discussion

3.1 Identification, concentration and content analysis of whole genomic DNA

A clear band of genome DNA size can be seen through the detection of 1% agarose gel electrophoresis. The detection value of OD260/OD280 by the microspectrophotometer is 1.8–1.9, which indicates that the purity of the DNA sample is better to perform the PCR experiment and library construction. The information of chloroplast clean reads, obtained by filtering it through fastq and the reads were compared to chloroplast reference.

3.2 Features of *C. spicatus* chloroplast genome

The chloroplast genome of *C. spicatus* is 152155bp, including a large single copy (LSC) region of 83098bp, small single copy (SSC) region of 17665bp, and a pair of inverted repeat regions (IRa and IRb) of 25696bp by each (Fig. 1). The GC content in the whole chloroplast genome of *C. spicatus* is 37.86%

and that of LSC, SSC and IR areas are 35.90%, 31.75% and 43.12%, respectively (Table 1). The GC content of the IR region is higher than that of the SSC region and the LSC region.

Table 1
Base composition in the chloroplast genome of *Clerodendranthus spicatus*

Region	Total	A (bp)	T (bp)	C (bp)	G (bp)	A + T (bp)	C + G (bp)	GC content(%)
LSC	83098	26002	27263	15257	14575	53265	29832	35.90
SSC	17665	6034	6021	2677	2932	12055	5609	31.75
IRb	25696	7324	7292	5337	5744	14616	11081	43.12
IRa	25696	7292	7324	5744	5337	14616	11081	43.12
Total	152155	46652	47900	29015	28588	94552	57603	37.86

3.3 Gene Content

One hundred and thirty-one genes in the circular genome of *C. spicatus*, including 87 protein-coding genes, 36 tRNA genes, and 8 rRNA genes were successfully annotated (Table 2). Fifteen protein-coding genes (*rps*16, *rps*7, *rpl*2, *rpl*23, *ndh*A, *ndh*B, *ndh*D, *ndh*E, *ndh*G, *ndh*H, *ndh*I, *psa*C, *ycf*1, *ycf*15 and *ycf*2), 7 tRNA genes (*trn*K-UUU, *trn*T-CGU, *trn*L-UAA, *trn*E-UUC (×2), *trn*A-UGC (×2)), and 8 rRNA genes (*rrn*16S (×2), *rrn*23S (×2), *rrn*4.5S (×2), *rrn*5S (×2)) are located in the IR region. Among these genes, nineteen cis-splicing genes, contain one or two intron, e.g., ten CDS (*rps*16(×2), *atp*F, *rpo*C1, *pet*D, *rpl*2(×2), *ndh*B(×2) and *ndh*A) and 7 tRNA genes contain one intron two kinds of protein-coding genes (*ycf*3 and *clp*P) contain two introns and three exons. Furthermore, the *rps*12 is a trans-splicing gene, which also contain three exons. As shown in the Fig. 1.. The white areas present introns, and the black areas stand for exons (Table 2 and Fig. 2, 3). The structures of trans-splicing genes in CDS from the plastome of *C. spicatus* are shown in Fig. 4. The white area is exon 2 in IRa, the black area is antother the exon 2 in IRb and the grey area is the exon1 (Fig. 4). The arrow shows the sense direction of the forward and reverse genes.

Table 2
Genes Contents of the chloroplast genome of *Clerodendranthus spicatus*

Category for genes	Group of genes	Name of genes
rRNA	rRNA genes	*rrn*16S (×2), *rrn*23S (×2), *rrn*5S (×2), *rrn*4.5S (×2)
tRNA	tRNA genes	36 unique trna genes (6 contains 1 intron)
Self-replication	Small subunit of ribosome	*rps*11, *rps*12(×2), *rps*14, *rps*15, *rps*16, *rps*18, *rps*19, *rps*2, *rps*3, *rps*4, *rps*7(×2), *rps*8
	Large subunit of ribosome	*rpl*14, *rpl*16, *rpl*2(×2), *rpl*20, *rpl*22, *rpl*23(×2), *rpl*32, *rpl*33, *rpl*36
	DNA dependent RNA polymerase	*rpo*A, *rpo*B, *rpo*C1, *rpo*C2
Photosynthesis	Subunits of NADH-dehydrogenase	*ndh*A, *ndh*B (×2), *ndh*C, *ndh*D, *ndh*E, *ndh*F, *ndh*G, *ndh*H, *ndh*I, *ndh*J, *ndh*K
	Subunits of photosystem I	*psa*A, *psa*B, *psa*C, *psa*I, *psa*J
	Subunits of photosystem II	*psb*A, *psb*B, *psb*C, *psb*D, *psb*E, *psb*F, *psb*I, *psb*J, *psb*K, *psb*L, *psb*M, *psb*N, *psb*T, *psb*Z, *ycf*3
	Subunits of cytochrome b/f complex	*pet*A, *pet*B, *pet*D, *pet*G, *pet*L, *pet*N
	Subunits of ATP synthase	*atp*A, *atp*B, *atp*E, *atp*F, *atp*H, *atp*I
	Large subunit of rubisco	*rbc*L
Other genes	Maturase	*mat*K
	Protease	*clp*P
	Envelope membrane protein	*cem*A
	Subunit of Acetyl-CoA-carboxylase	*acc*D
	c-type cytochrom synthesis gene	*ccs*A
Genes of unknown functions		*ycf*1, *ycf*15 (×2), *ycf*2 (×2), *ycf*4
The “(x2)” indicates that the gene located in the IRs and thus had two copies.		

Table 3
The lengths of introns and exons in genes from chloroplast genome of *Clerodendranthus spicatus*

Gene	Location	Start	End	length(bp)				
				Exon I	Intron I	Exon II	Intron II	Exon III
*trn*K-UUU	LSC	1672	4250	37	2506	36		
*rps*16	LSC	4777	5910	40	867	227		
*trn*T-CGU	LSC	8970	9733	35	686	43		
atpF	LSC	11725	12965	145	686	410		
rpoC1	LSC	20741	23547	432	764	1611		
ycf3	**LSC**	**41971**	**43919**	**126**	**710**	**228**	**732**	**153**
trnL-UAA	LSC	46897	47453	35	472	50		
clpP	**LSC**	**69239**	**71170**	**71**	**701**	**294**	**640**	**226**
petD	LSC	75665	76848	8	701	475		
rpl16	LSC	80283	81511	9	821	399		
rpl2	IRb	83203	84681	391	654	434		
ndhB	IRb	93392	95599	775	675	758		
rps12	**IRb**	**96927**	**97169**	**242**	**-**	**25**	**357**	**113**
trnE-UUC	IRb	100907	101924	32	946	40		
trnA-UGC	IRb	101989	102866	37	805	36		
ndhA	SSC	115271	117373	553	1011	539		
trnA-UGC	IRa	132388	133265	37	805	36		
trnE-UUC	IRa	133330	134347	32	946	40		
rps12	**IRa**	**138085**	**138327**	**113**	**-**	**242**	**357**	**25**
ndhB	IRa	139655	141862	775	675	758		
rpl2	IRa	150573	152051	391	654	434		

3.4 The characteristics of rRNAs and tRNAs genes

There are 8 rRNA genes in the chloroplast genome of *C. spicatus*, including *rrn*16S (×2), *rrn*23S (×2), *rrn*4.5S (×2), *rrn*5S (×2) with the inverse direction by one pairs. The length of *rrn*16S, *rrn*23S, *rrn*4.5S and *rrn*5S is 1491bp, 2811bp, 265bp and 131bp, respectively. The GC contents of them are 56.54%, 54.93%,45.28% and 50.38%, respectively (Fig. 5). The average of GC contents is 51.78%. Through the

scanning of tRNAs, we can find 18 types of amino acids can be transported, including Arg, Asn, Asp, Cys, Gly, Gln, Glu, His, Ile, Leu, Met, Phe, Pro, Ser, Trp, Thr, Tyr and Val, of which the anti-codons are TCT and ACG, GTT, GTC, GCA, GCC, TTG, TTC, GTG, GAT, CAA and TAG, CAT, GAA, TGG, TGA, GGA and GCT, CCA, GGC and TGT, GTA, GAC, respectively (Table 4).

Table 4
The scanned tRNA from chloroplast genome of *Clerodendranthus spicatus*

tRNA No.	tRNA Bounds		tRNA Type	Anti-Codon	Intron Bounds		Score
	Begin	End			Begin	End	
1	9924	9995	Arg	TCT	0	0	67.2
2	27973	28043	Cys	GCA	0	0	60.7
3	30965	31036	Thr	GGT	0	0	67.4
4	35835	35905	Gly	GCC	0	0	61.4
5	44762	44848	Ser	GGA	0	0	72.2
6	47736	47808	Phe	GAA	0	0	71.9
7	51851	51923	Met	CAT	0	0	59.5
8	98827	98898	Val	GAC	0	0	58.9
9	100907	100994	Ile	GAT	100943	100958	16.4
10	106662	106735	Arg	ACG	0	0	57.7
11	127874	127945	Asn	GTT	0	0	72.5
12	142432	142512	Leu	CAA	0	0	62.3
13	150034	150107	Met	CAT	0	0	70.3
14	136427	136356	Val	GAC	0	0	58.9
15	134347	134260	Ile	GAT	134311	134296	16.4
16	128592	128519	Arg	ACG	0	0	57.7
17	122867	122788	Leu	TAG	0	0	58.5
18	107380	107309	Asn	GTT	0	0	72.5
19	92822	92742	Leu	CAA	0	0	62.3
20	85220	85147	Met	CAT	0	0	70.3
21	66129	66056	Pro	TGG	0	0	65.1
22	65887	65814	Trp	CCA	0	0	71.9
23	46208	46136	Thr	TGT	0	0	69.3
24	36151	36078	Met	CAT	0	0	63.2
25	35068	34976	Ser	TGA	0	0	78.4
26	30363	30291	Glu	TTC	0	0	56.2

tRNA No.	tRNA Bounds		tRNA Type	Anti-Codon	Intron Bounds		Score
	Begin	End			Begin	End	
27	30207	30124	Tyr	GTA	0	0	62.3
28	30006	29933	Asp	GTC	0	0	67.9
29	8231	8144	Ser	GCT	0	0	74.5
30	6922	6851	Gln	TTG	0	0	58.9
31	85	12	His	GTG	0	0	60.7

3.5 Analysis of RSCU and CAI regarding the codon usage

Gene sequence and the frequency of genetic code usages are closely related to plant evolution and genetic relationship [38]. From the codon usage analysis of results from the chloroplast genome of *C. spicatus*, there were 26421 codons in all protein-coding genes, of which Isoleucine (1102 codons, accounting for 4.17% of the whole codons) were the richest amino acid in the *C. spicatus*. chloroplast genomes [39]. Lysine was the second richest amino acid, accounting for 4.06% of the whole codons, while cysteine only had 0.24% of the whole codons. Besides, total of the fractions and frequencies for all codons usage are 21.001 and 1000, respectively (Table 5). Therefore it can be found that different codon appear and use in different frequency and RSCU values are an indication of how many times the codon is observed relative to the number of times it should be observed in the absence of any codon usage bias for a particular amino acid related to the evolutionary of species [40] (Table 5). The RSCU varied from 0.336 to 2.9904. RSCU value > 1 for each codon shows that this codon is preferred. In this study, the codons of AUG, UUA and AGA codon had the higher RSCU value indicating the presence of higher codon usage bias in total of 61 genes. The CAI value is 0.645 indicating the codon preference of genes. Excluding the stop codons, only two amino acids (Trp and Met) are encoded by a kind of codon, respectively [41].

Table 5
Codon usage and codon-anticodon recognition patterns in *Clerodendranthus spicatus* chloroplast genome

AminoAcid	Symbol	Codon	No.	Fraction	Frequency	RSCU	tRNA
A	Ala	GCA	385	0.271	13.797	1.0944	*trn*A-UGC
A	Ala	GCC	224	0.151	7.68	0.6368	-
A	Ala	GCG	171	0.123	6.277	0.486	-
A	**Ala**	**GCU**	**627**	**0.454**	**23.081**	**1.7824**	**-**
C	**Cys**	**UGC**	**64**	**0.261**	**3.068**	**0.4354**	trn**C-GCA**
C	**Cys**	**UGU**	**230**	**0.739**	**8.703**	**1.5646**	**-**
D	Asp	GAC	210	0.199	8.081	0.3922	*trn*D-GUC
D	**Asp**	**GAU**	**861**	**0.801**	**32.627**	**1.6078**	**-**
E	**Glu**	**GAA**	**1032**	**0.754**	**38.803**	**1.5166**	trnE-UUC
E	Glu	GAG	329	0.246	12.634	0.4834	-
F	Phe	UUC	499	0.337	19.532	0.6712	*trnF-GAA*
F	Phe	UUU	988	0.663	38.442	1.3288	-
G	**Gly**	**GGA**	**725**	**0.391**	**25.187**	**1.6264**	trnG-UCC
G	Gly	GGC	204	0.126	8.142	0.4576	*trnG-GCC*
G	Gly	GGG	313	0.188	12.092	0.702	-
G	Gly	GGU	541	0.296	19.071	1.2136	-
H	His	CAC	141	0.245	5.936	0.4608	*trnH-GUG*
H	His	CAU	471	0.755	18.249	1.5392	-
I	Ile	AUA	680	0.303	25.147	0.9084	-
I	Ile	AUC	464	0.206	17.126	0.6198	*trnI-GAU*
I	**Ile**	**AUU**	**1102**	**0.491**	**40.849**	**1.4721**	**-**
K	**Lys**	**AAA**	**1072**	**0.727**	**40.849**	**1.4816**	trnK-UUU
K	Lys	AAG	375	0.273	15.361	0.5184	-
L	Leu	CUA	397	0.132	13.977	0.8412	*trnL-UAG*
L	Leu	CUC	185	0.066	6.999	0.3918	-
L	Leu	CUG	194	0.073	7.68	0.411	-

AminoAcid	Symbol	Codon	No.	Fraction	Frequency	RSCU	tRNA
L	Leu	CUU	605	0.214	22.54	1.2822	-
L	Leu	UUA	869	0.3	31.644	**1.842**	*trnL-UAA*
L	Leu	UUG	581	0.215	22.68	1.2312	*trnL-CAA*
M	**Met**	**AUG**	**620**	**1**	**22.54**	**2.9904**	trnI-GAU
N	Asn	AAC	302	0.247	11.811	0.479	*trnN-GUU*
N	Asn	AAU	959	0.753	36.016	1.521	-
P	Pro	CCA	308	0.283	11.711	1.11	*trnP-UGG*
P	Pro	CCC	229	0.215	8.864	0.8252	-
P	Pro	CCG	165	0.159	6.557	0.5944	-
P	Pro	CCU	408	0.343	14.178	1.4704	-
Q	Gln	CAA	711	0.756	27.112	1.5242	*trnQ-UUG*
Q	Gln	CAG	222	0.244	8.763	0.4758	-
R	Arg	AGA	495	0.299	18.589	**1.8264**	*trnR-UCU*
R	Arg	AGG	172	0.119	7.42	0.6348	-
R	Arg	CGA	367	0.218	13.516	1.3542	-
R	Arg	CGC	120	0.074	4.572	0.4428	-
R	Arg	CGG	130	0.084	5.194	0.48	-
R	Arg	CGU	342	0.206	12.814	1.2618	*trnR-ACG*
S	Ser	AGC	116	0.061	4.933	**0.336**	*trnS-GCU*
S	Ser	AGU	421	0.202	16.203	1.2198	-
S	Ser	UCA	401	0.189	15.16	1.1616	*trnS-UGA*
S	Ser	UCC	351	0.168	13.456	1.017	*trnS-GGA*
S	Ser	UCG	197	0.099	7.921	0.5706	-
S	Ser	UCU	585	0.282	22.6	1.695	-
T	Thr	ACA	397	0.292	14.358	1.194	*trnT-UGU*
T	Thr	ACC	247	0.188	9.225	0.7428	*trnT-GGU*
T	Thr	ACG	144	0.112	5.515	0.4332	-
T	Thr	ACU	542	0.408	20.053	1.63	-

AminoAcid	Symbol	Codon	No.	Fraction	Frequency	RSCU	tRNA
V	Val	GUA	541	0.37	19.472	1.5196	*trnV-UAC*
V	Val	GUC	172	0.127	6.698	0.4832	*trnV-GAC*
V	Val	GUG	181	0.133	6.999	0.5084	-
V	Val	GUU	530	0.369	19.432	1.4888	-
W	Trp	UGG	465	1	18.469	1	*trnW-CCA*
Y	Tyr	UAC	186	0.197	7.119	0.3896	*trnY-GUA*
Y	Tyr	UAU	769	0.803	28.937	1.6104	-
Stop	Ter	UAA	46	0.423	3.188	1.5861	-
Stop	Ter	UAG	25	0.309	2.326	0.8622	-
Stop	Ter	UGA	16	0.269	2.025	0.5517	-

3.6 Repeat Sequences analysis

Repeat sequences are kinds of important genetic markers and are closely related to the origin and evolution of species [42]. Repeat sequences can generally be divided into scattered(interspersed) repetition and tandem repetition sequence (TRS) [43]. Interspersed repetition sequences are scattered in the way and distributed in the genome. Multiple repeats of a sequence on a chromosome are called tandem repeats. A special form of tandem repeats is simple tandem repeats, also known as simple repeats (SSR) [44]. SSRs often have natural polymorphism. The cpSSRs are the focus of chloroplast genome repeat analysis. Therefore, we analyzed three types of repetitive sequences (simple sequence repeats, tandem repeats, and interspersed repeats) in the chloroplast genome of *C. spicatus*. For the simple sequence repeat, 28 repeats (12 A,15T and 1 TA) were identified and the types are P1and P2 (Table S2). The sizes of SSRs are between 10bp and 15bp (Table S2). For the tandem repeats, 36 repeats were identified in the chloroplast genome of *C. spicatus*, which conformed with the two conditions that the length of the repeat unit was more than 20 bp and the similarity among the repeat unit sequences was more than 70% (Table S3) [45]. For interspersed repeats, 40 repeats were identified including 22 palindromic repeats and 18 direct repeats (Table S4). The length of repeat unit 1, 2 are between 30bp and 60bp. The E-values of interspersed repeats in the chloroplast genome of *C. spicatus* are from 7.80E-23 to 6.19E-04 (Table S4).

3.7 IR structures and genes analysis of seven selected species

The sizes of the four regions of the chloroplast genome from 7 selected species were analyzed, while the boundary between each two adjacent regions were also analyzed. The results showed that the selected chloroplast genomes had the diverse similiar structures confirming with the different sizes of four areas (Fig. 6). For the seven species, the *rps*19 genes were located in the border area of LSC and IRb in the species of *C. spicatus, Salvia miltiorrhiza, Salvia miltiorrhiza f. alba* and *Tectona grandis*. However, the

rps19 genes were located in the area of LSC from *Epimedium brevicornu* and located in the area of IRb area from *Cistanche deserticola.* The *rps*19 gene had unstable position relationship, sometimes across the area of LSC-IRb, while other times only within the areas of LSC or IRb. There is no *rps*19 and *ycf*1 genes found in *Glechoma longituba*. In contrast, *ndhF* genes were located at the border area of IRb and SSC in the species of *Salvia miltiorrhiza, Salvia miltiorrhiza f. alba* and *Tectona grandis* and located at area of SSC in the species of *Epimedium brevicornu* [46]. Most of the *ycf*1 genes were located in the border area of SSC and IRa in the genus of *Salvia* and *Tectona grandis*,while it was also located in the area of SSC and IRb in *C. spicatus, Epimedium brevicornu,Glechoma longituba* and *Tectona grandis*. Besides, *rpl*2, *trn*H gene spacers were located in the IRa and LSC areas respectively, except the species of *Glechoma longituba*, respectively. A small fragment of the *rps*19 gene was found in the border area of IRa and LSC in the genus of *Salvia* and *Cistanche deserticola*. There had the *psb*A genes existing in the genus of *Salvia*, *Tectona grandis and C. spicatus.*

3.8 Hypervariable region identification

To investigate the chloroplast genome divergence of 5 relative species, those are *C. spicatus, Salvia miltiorrhiza, Salvia miltiorrhiza f. alba, Tectona grandis,Glechoma longituba* based on the results of phylogenetic analysis, we conducted a genetic distance analysis of intergenic spacer regions (IGS) for them. The result showed 30 out of 58 intergenic spacer regions were identical with K2p (Kimura 2-parameter) values varying from 6.084 to 29.242 (Fig. 7), of which five intergenic spacer regions had higher K2p values and variation above the value of 18.0, namely, *ndh*G-*ndh*I (29.242), *acc*D-*psa*I (22.442), *rps*15-*ycf*1 (19.488), *rpl*20-*clp*P (18.322), and *ccs*A-*ndh*D (18.091). We can develop the specific molecular markers within the variations of these IGS regions and use them to distinctively identify the species [47].

3.9 Phylogenetic analysis

The structure of chloroplast genome is simple and the length is small. The sequence of it is conserved and genes are mostly orthologous. Therefore it is of great value to study the evolution relationship between green plants and the chloroplast genome. In this study, *Clerodendranthus spicatus* is from a kind of single genus from Lamiaceae family. There are 82 CDS shared gene nucleic acid sequences were extracted from the 15 species and used to construct the phylogenetic trees (Table S5). Among the CDS, some species,are distinct with the proteins after comparisons. *Cistanche deserticola* is special in the proteins of *psbM, rpl*14, *rpl*33, *rpl*36, *rps*3, *rps*4, *rps*7, *rps*12, *rps*14, *rps*16, *rps*18, *rps*19 and *ycf*4. In addition, the proteins of *psb*Z, *rpo*C2 are common in the 13 species. However, *psb*Z is loss in the species of *Cistanche deserticola* and *Glechoma longituba. rpo*C2 is loss in *Cistanche deserticola* and *Tectona grandi.* The protein of *ycf*15 exists in 14 species except the *Epimedium brevicornu.* The proteins of *rpo*C and *lhb*A are only common in the species of *Tectona grandis, Glechoma longituba*, apartly. In the species of *Dipsacus asper, proteins of ndh*I, *psb*C, *rpl*22, *rpo*B, *rps*2, *rps*14, *rps*18 *are diverse from others.* The *psb*I *in Eucommia ulmoides, rps*3 *in Rheum tanguticum, rps*12 *and ycf*4 *in Cullen corylifolium are distinct* (Table S5).

The phylogenetic tree showed that 13 species are clustered into one branch except two outgroups, the species of *Epimedium brevicornu* and *Dioscorea polystachya.* Bootstrap analysis showed that there were 7 out of 11 nodes with 100% bootstrap values. At the one branch, the tree is subdivided into two branches, the 6 species of *Lamiales*, *Eucommia ulmoides* and *Dipsacus asper* are clustered together, otherwise, 4 species from *Polygonaceae family and Cullen corylifolium* from *Fabaceae* family were clustered together (Fig. 8) [48]. It indicated that the herbaceous plants *C. spicatus*, two genus of *Salvia and Glechoma longituba* from *Lamiaceae* family were closely related genetic relationship. The xylophyta species of *Tectona grandis* is some correlation to these four plants above with bootstrap value of 87. *Cistanche deserticola* is clustered into single position within the one branch because it maybe a parasitic plant parasitic at the roots of the tree shuttle in the desert differed from others about the many proteins and specific genes. Nevertheless, the two plants of *Epimedium brevicornu* and *Dioscorea polystachya* were not related to the *C. spicatus* and were more distant relationships with others.

Conclusions

We finished the sequencing, assembly and annotation of *C. spicatus* chloroplast genome based on the next-generation sequencing technology. The unique genetic characteristics could provide material resources for the subsequent genetic analysis. Sole genus of *C. spicatus* is genetically closer to the genus of *Salvia* and *Glechoma longituba*. Therefore, the chloroplast genome information mining of *C. spicatus* has well complemented the phylogenetic relationships of the Lamiaceae family, which has provided the worthy data for a further understanding of the genetic development of this family. Regions of *ndh*G-*ndh*I, *acc*D-*psa*I, *rps*15-*ycf*1, *rpl*20-*clp*P, *ccs*A-*ndh*D within 58 IGS tend to demonstrate higher genetic polymorphism. These genes can be used as a kind of molecular marker for subsequent application. The results strongly provide the valuable information to study the development and evolutionary correlation of single species with similar effects on treatment of the renal diseases. In the future, we hope to combine germplasm gene resources to study the transcripts, produce proteins of genes in chloroplast genome to provide the valuable data for the production,, and change the functional application of diverse chemical substances.

Abbreviations

LSC: large single copy

SSC: small single copy

IR: inverted repeat

SSR: Simple Sequence Repeat

ISSR: inter-simple sequence repeat

cpSSR: chloroplast simple sequence repeat

UPGM A: unweighted pair-group method with arithmetic means

DNA: DeoxyriboNucleic Acid

tRNA: Transfer RNA

rRNA: ribosomal RNA

RSCU:Relative Synonymous Codon Usage

CAI: Codon Adaptation Index

TRF: Tendem Repeat Finder

MISA: microsatellites

EMBOSS: European Molecular Biology Open Software Suite

K2p: Kimura 2-parameters

CDS: Coding sequence

ML: maximum likelihood

MEGA: molecular evolutionary genetics analysis

OD: Optical Density

PCR: polymerase chain reaction

GC: Guanine and cytosine

IGS: intergenic spacer regions

Declarations

Acknowledgements

I would like to extend my deep gratitude to Dr. Niyang, Dr. Li Jingsheng, Miss Yue Jingwen and Mr. Zhou Junchen who have offered mepractical, cordial and selfless support in analyzing the data of manuscript.

Author Contribution

CL conceived the study; MJ and LQW collected samples of *Clerodendranthus spicatus*, extracted DNA for next-generation sequencing; DQ, SSH, SYL assembled, validated the genome, performed data analysis and drafted the manuscript; HMC reviewed the manuscript critically; CBJ and HDG reviewed and put forward the revised advice. All authors have read and agreed the contents of the manuscript.

Funding

This work was supported by the National Science &Technology Fundamental Resources Investigation Program of China [2018FY100705], The National Mega-Project for Innovative Drugs of China [2019ZX09735-002], Chinese Academy of Medical Sciences, Innovation Funds for Medical Sciences (CIFMS) [2016-I2M-3-016, 2017-I2M-1-013], National Science Foundation Funds [81872966], Key Laboratory of Medicinal Animal and Plant Resources of Qinghai-Tibetan Plateau in Qinghai Province [2020-ZJ-40], Qinghai Provincial Key Laboratory of Phytochemistry of Qinghai Tibet Plateau[2017-ZJ-Y20]. The funders were not involved in the study design, data collection, analysis, decision to publish, or manuscript preparation.

Compliance with ethical standards

Conflict of interest All the authors declare no conflicts of interest.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

Data availability statement

The genome sequence data of *Clerodendranthus spicatus* that support the findings of this study are openly available in NCBI at GenBank database with accession number MZ063774. (https://www.ncbi.nlm.nih.gov). The associated BioProject, SRA, and Bio-Sample numbers are PRJNA723363, SRR14350365 and SAMN18814878, respectively.

References

1. Acta Phytotax (1974) Sin.,12:233.
2. Xie, Z.Y (1996) National Herbal Compendium, 2nd ed.; Kuang, L.J., Ed.; People's Medical Publishing House: Beijing, China, p.572.
3. Jia, M.R.; Li, X.W (2005) Chinese Ethnic Materia Medica, 1st ed.; Zhao, Y.X., Ed.; China Medical Science and Technology Press: Beijing, China, pp 166.
4. Zhao A.H., Zhao Q.S., Li R.T., Sun H.D (2004) Chemical constituents from *Clerodendranthus spicatus*. Acta Bot. Yunnanica 26:563-568.
5. Qingxia Zheng, Zhaocui Sun, Xiaopo Zhang, Jinquan Yuan, Haifeng Wu, Junshan Yang and Xudong Xu (2012) Clerodendranoic Acid, a New Phenolic Acid from *Clerodendranthus spicatus*. Molecules 17(11):13656-13661.
6. Wang M, Liang J.Y., Chen X.Y (2007) Water-soluble constituents of *Clerodendranthus spicatus*. Chin. J. Nat. Med 5:27-30.
7. Chen Y.L., Tan C.H., Tan J.J., Zhao X.M., Jiang S.H., Zhu D.Y (2009) Two new diterpenoid glucosides from *Clerodendranthus spicatus*. Helv. Chim. Acta 92:2802-2807.

https://doi.org/10.1002/hlca.200900121.

8. Zhong JY, Wu ZS (1984) Chemical constituents of *Clerodendranthus spicatus***.** Acta Bot. Yunnanica 6(3):344-345.
9. Zou J, Zhu Y.D., Zhao W.M (2008) Two new alkyl glycosides from *Clerodendranthus spicatus*. J. Asian Nat. Prod. Res 10(7-8):603-607.doi:10.1080/10286020802133076.
10. Zhu Chen, Caixia Yang, Xiaojian Mao (2015) Progress in Related Pharmacological Research on Function of *Clerodendranthus spicatus*. Pharmacy Information 4(2):23-30. doi:10.12677/PI.2015.42003.
11. Zhang Yongyi, Wu Jiachao, Li Shuiping, Li Nili, Huang Yongfei, Lin Qinghua (2021) Research progress on chemical constithents and pharmacological actions of *Clerodendruntius spicatus*. Acta Chinese Medicine and Pharmacology 49 (1):112-120.
12. Pedro Robles and Víctor Quesada. Organelle Genetics in Plants (2021) Int J Mol Sci 22(4):2104. doi:10.3390/ijms22042104.
13. Sidsel Birkelund Schmidt, Marion Eisenhut, Anja Schneider (2020) Chloroplast Transition Metal Regulation for Efficient Photosynthesis. Trends Plant Sci 25(8):817-828.doi: 10.1016/j.tplants.2020.03.003.
14. Judith Van Dingenen, Jonas Blomme, Nathalie Gonzalez and Dirk Inzé (2016) Plants grow with a little help from their organelle friends. J Exp Bot. 67(22):6267-6281.doi: 10.1093/jxb/erw399.
15. Ai-Sheng Xiong, Ri-He Peng, Jing Zhuang, Feng Gao, Bo Zhu, Xiao-Yan Fu, Yong Xue, Xiao-Feng Jin, Yong-Sheng Tian, Wei Zhao, Quan-Hong Yao (2009) Gene duplication, transfer, and evolution in the chloroplast genome. Biotechnol Adv. 27(4):340-347.doi: 10.1016/j.biotechadv.2009.01.012.
16. Gray M.W. Evolution of organellar genomes (1999) Curr. Opin. Genet. Dev. 9:678–687.doi: 10.1016/S0959-437X(99)00030-1.
17. Henry Daniell, Choun-Sea Lin, Ming Yu, Wan-Jung Chang (2016) Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biol. 17(1):134. doi:10.1186/s13059-016-1004-2.
18. Luo Can, Yue xudong, Wu Fanhua, Gao Xin, Wang Tongxin (2015) Genetic diversity in *Clerodendranthus spicatus* (Thunb.) based on ISSR markers. Journal of troptical biology 6(2): 204-222.doi:10.15886/j.cnki.rdswxb.2015.02.016.
19. Deepti Nigam and Hernan Garcia-Ruiz (2020) Variation Profile of the Orthotospovirus Genome. Pathogens 9(7):521.doi: 10.3390/pathogens9070521.
20. Catherine J. Nock, Daniel L.E. Waters, Mark A. Edwards, Stirling G. Bowen, Nicole Rice, Giovanni M. Cordeiro and Robert J Henry (2011) Chloroplast genome sequences from total DNA for plant identification. Plant Biotechnol J 9(3):328-333.doi: 10.1111/j.1467-7652.2010.00558.
21. Cronn R., Liston A., Parks M., Gernandt D., Shen R. and Mockler T (2008). Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res.36(19): e122.doi: 10.1093/nar/gkn502.

22. Anthony M Bolger, Marc Lohse, Bjoern Usadel (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15): 2114-2120. doi: 10.1093/ bioinformatics /btu170.
23. Dierckxsens N., Mardulyn P. and Smits G(2016) NOVOPlasty:De novo assembly of organelle genomes from whole genome data. Nucleic Acids Research 45(4): e18.doi:10.1093/nar/gkw955.
24. Linchun Shi, Haimei Chen, Mei Jiang, Liqiang Wang, Xi Wu, Linfang Huang,Chang Liu (2019) CPGAVAS2, an integrated plastome sequence annotator and analyzer. Nucleic Acids Research 47(W1): 65-73.doi: 10.1093/nar/gkz345.
25. Can Firtina, Jeremie S. Kim, Mohammed Alser, Damla Senol Cali, A. Ercument Cicek., Can Alkan, and Onur Mutlu (2020) Apollo: a sequencing-technology-independent, scalable and accurate assembly polishing algorithm. Bioinformatics 36(12):3669-3679.doi: 10.1093/bioinformatics/btaa179.
26. Peden J.F (1999) Analysis of codon usage.PhD Thesis, University of Nottingham, UK.
27. Puigbo P, Bravo I. G. and Garcia-Vallve, S (2008) CAIcal: a combined set of tools to assess codon usage adaptation. Biol. Direct. 3:38.doi:10.1186/1745-6150-3-38.
28. Stefan Kurtz,a Jomuna V. Choudhuri, Enno Ohlebusch, Chris Schleiermacher, Jens Stoye, and Robert Giegerich (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res 29(22):4633-4642. doi:10.1093/nar/29.22.4633.
29. G. Benson (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research 27(2):573-580.doi:10.1093/nar/27.2.573.
30. Beier S, Thiel T, Münch T, Scholz U, Mascher M (2017) MISA-web: a web server for microsatellite prediction. Bioinformatics, 33(16):2583–2585. https://doi.org/10.1093/bioinformatics/btx198.
31. Ali Amiryousefi1, Jaakko Hyvo nen and Peter Poczai (2018) IRscope: an online program to visualize the junction sites of chloroplast genomes. Bioinformatics 34(17):3030-3031. doi:10.1093/bioinformatics/bty220.
32. Rice P, Longden I, Bleasby A (2000) EMBOSS:the European Molecular Biology Open Software Suite. Trends Genet. 16(6):276-277.doi:10.1016/s0168-9525(00)02024-2..
33. Asim Kumar Mahadani, Shashank Awasthi, Goutam Sanyal, Partha Bhattacharjee and Sanjeev Pippal (2021) Indel-K2P: a modified Kimura 2 Parameters (K2P) model to incorporate insertion and deletion (Indel) information in phylogenetic analysis. Cyber-Physical Systems 7(1):1-13. https://doi.org/10.1080/23335777.2021.1879274.
34. Zhang D., F. Gao, I. Jakovlić, H. Zou, J. Zhang, W.X. Li, and G.T. Wang (2020) PhyloSuite: An integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. Molecular Ecology Resources 20(1):348–355.doi: 10.1111/1755-0998.13096.
35. Kazutaka Katoh, Daron M Standley (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30(4):772-780.doi: 10.1093/molbev/mst010.
36. Lam-Tung Nguyen, Heiko A Schmidt, Arndt von Haeseler, Bui Quang Minh (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol

32(1):268-274.doi: 10.1093/molbev/msu300.

37. Barry G Hall (2013) Building phylogenetic trees from molecular data with MEGA. Mol Biol Evol 30(5):1229-1235.doi: 10.1093/molbev/mst012.
38. Sharp P.M., Li W.H (1986) An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol 24:28–38. doi:10.1007/BF02099948.
39. Gustafsson C, Govindarajan S, Minshull J (2004) Codon bias and heterologous protein expression. Trends Biotechnol 22:346–353.
40. P M Sharp, W H Li (1986) An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 24(1-2):28-38. doi: 10.1007/BF02099948.
41. Fatemeh Chamani Mohasses, Mahmood Solouki, Behzad Ghareyazie ID, Leila Fahmideh, Motahhareh Mohsenpour (2020) Correlation between gene expression levels under drought stress and synonymous codon usage in rice plant by in-silico study. PLoS ONE 15(8):e0237334.https://doi.org/10.1371/journal.pone.0237334.
42. Jacques Nicolas (2012) To detect and analyze sequence repeats whatever be their origin. Methods Mol Biol 859:69-90.doi: 10.1007/978-1-61779-603-6_4.
43. Rafael Navajas-Pérez, Andrew H Paterson (2009) Patterns of tandem repetition in plant whole genome assemblies. Mol Genet Genomics 281(6):579-590.doi: 10.1007/s00438-009-0433-y.
44. Lin T.-Y (2016) Simple sequence repeat variations expedite phage divergence: Mechanisms of indels and gene mutations. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 789:48–56.doi: 10.1016/j.mrfmmm.2016.04.001.
45. Ellegren H (2004) Microsatellites: simple sequences with complex evolution. Nat Rev Genet 5:435–445.doi: 10.1038/nrg1348.
46. Jun Qian, Jingyuan Song, Huanhuan Gao, Yingjie Zhu, Jiang Xu, Xiaohui Pang, Hui Yao, Chao Sun, Xian'en Li, Chuyuan Li, Juyan Liu, Haibin Xu, Shilin Chen (2013) The Complete Chloroplast Genome Sequence of the Medicinal Plant Salvia miltiorrhiza. PLoS ONE 8(2): e57607. doi:10.1371/journal.pone.0057607.
47. Péter Poczai, Jaakko Hyvönen (2010) Nuclear ribosomal spacer regions in plant phylogenetics: problems and prospects. Mol Biol Rep 37(4):1897-1912.doi: 10.1007/s11033-009-9630-3.
48. Conglian Liang, Lei Wang, Juan Lei, Baozhong Duan f, Weisi Ma, Shuiming Xiao, Haijun Qi, Zhen Wang, Yaoqi Liu, Xiaofeng Shen, Shuai Guo, Haoyu Hu, Jiang Xu, Shilin Chen (2019) A Comparative Analysis of the Chloroplast Genomes of Four Salvia Medicinal Plants. Engineering 5:907-915.10.1016/j.eng.2019.01.017.

Figures

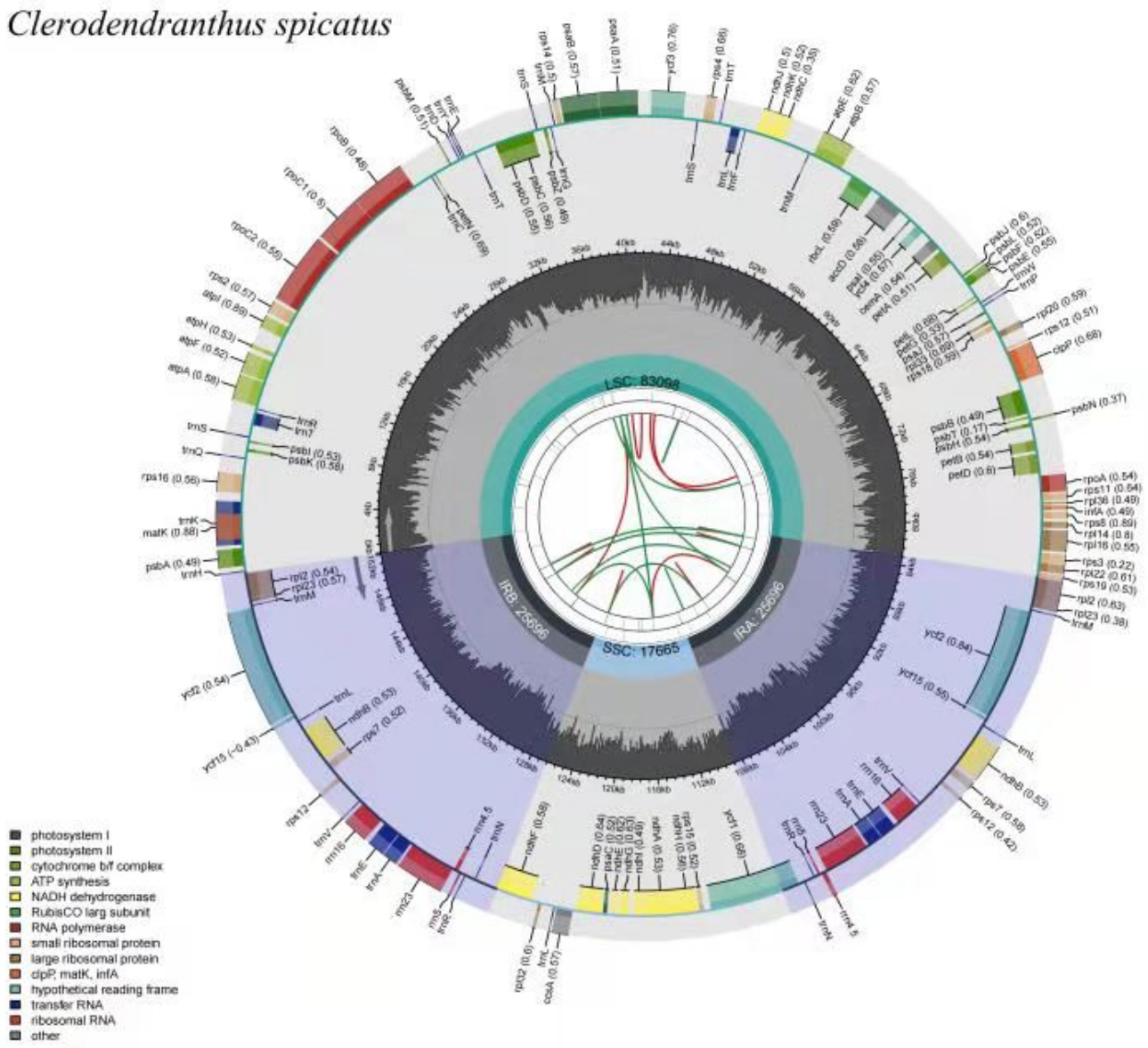

Figure 1

Graphic representation of features identified in Clerodendranthus spicatus chloroplast genome by using CPGview-RSG (http://www.herbalgenomics.org/cpgview). The map contains seven circles. From the center going outward, the first circle shows the distributed repeats connected with red (the forward direction) and green (the reverse direction) arcs. The next circle shows the tandem repeats marked with short bars. The third circle shows the microsatellite sequences as short bars. The fourth circle shows the size of the LSC and SSC. The fifth circle shows the IRA and IRB. The sixth circle shows the GC contents along the plastome. The seventh circle shows the genes having different colors based on their functional groups.

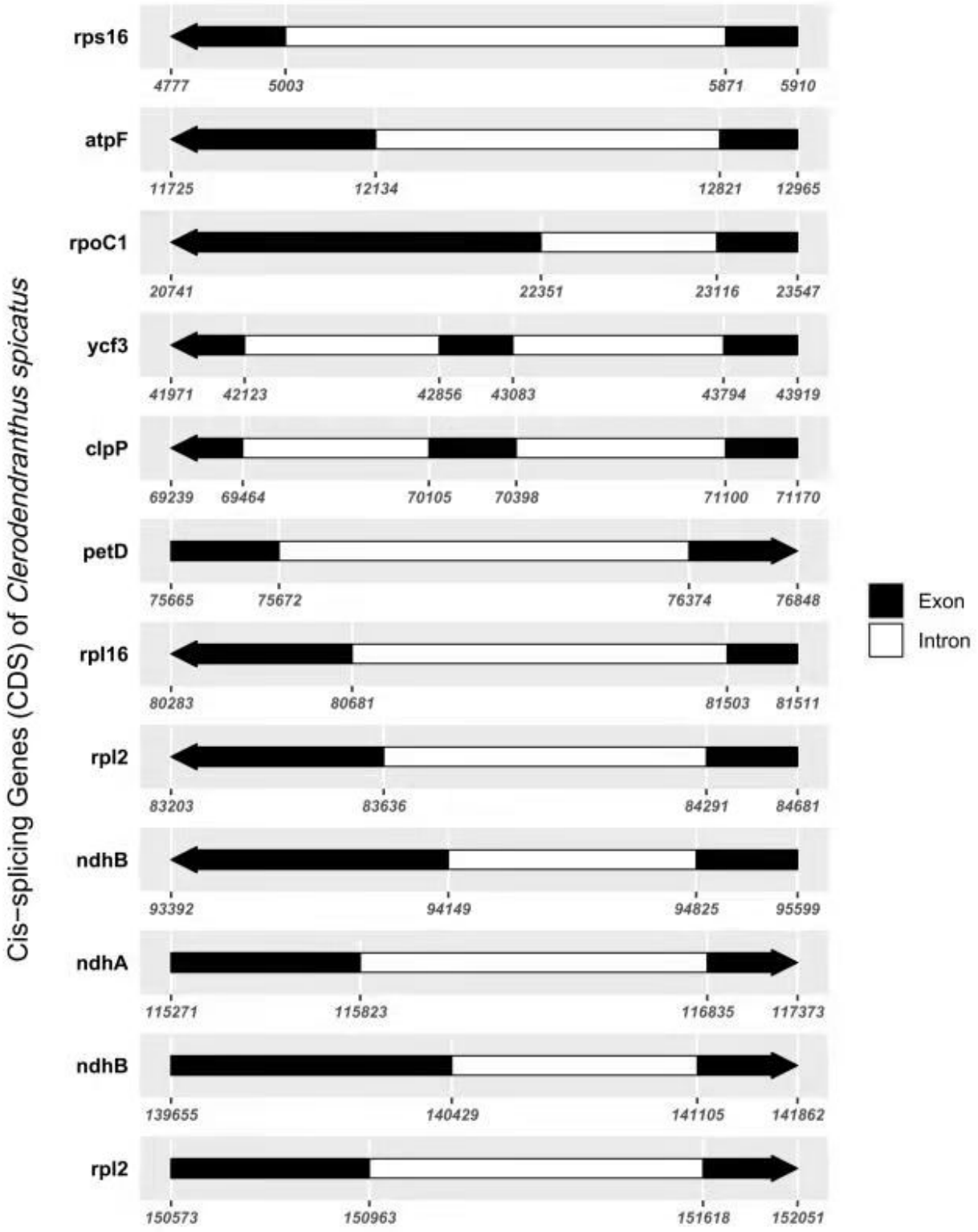

Figure 2

Schematic presentation of the structure of cis-splicing genes (CDS) from the plastome of Clerodendranthus spicatus (http://www.herbalgenomics.org/cpgview). The white areas present introns, and the black area present exons. The arrow shows the sense direction of the genes.

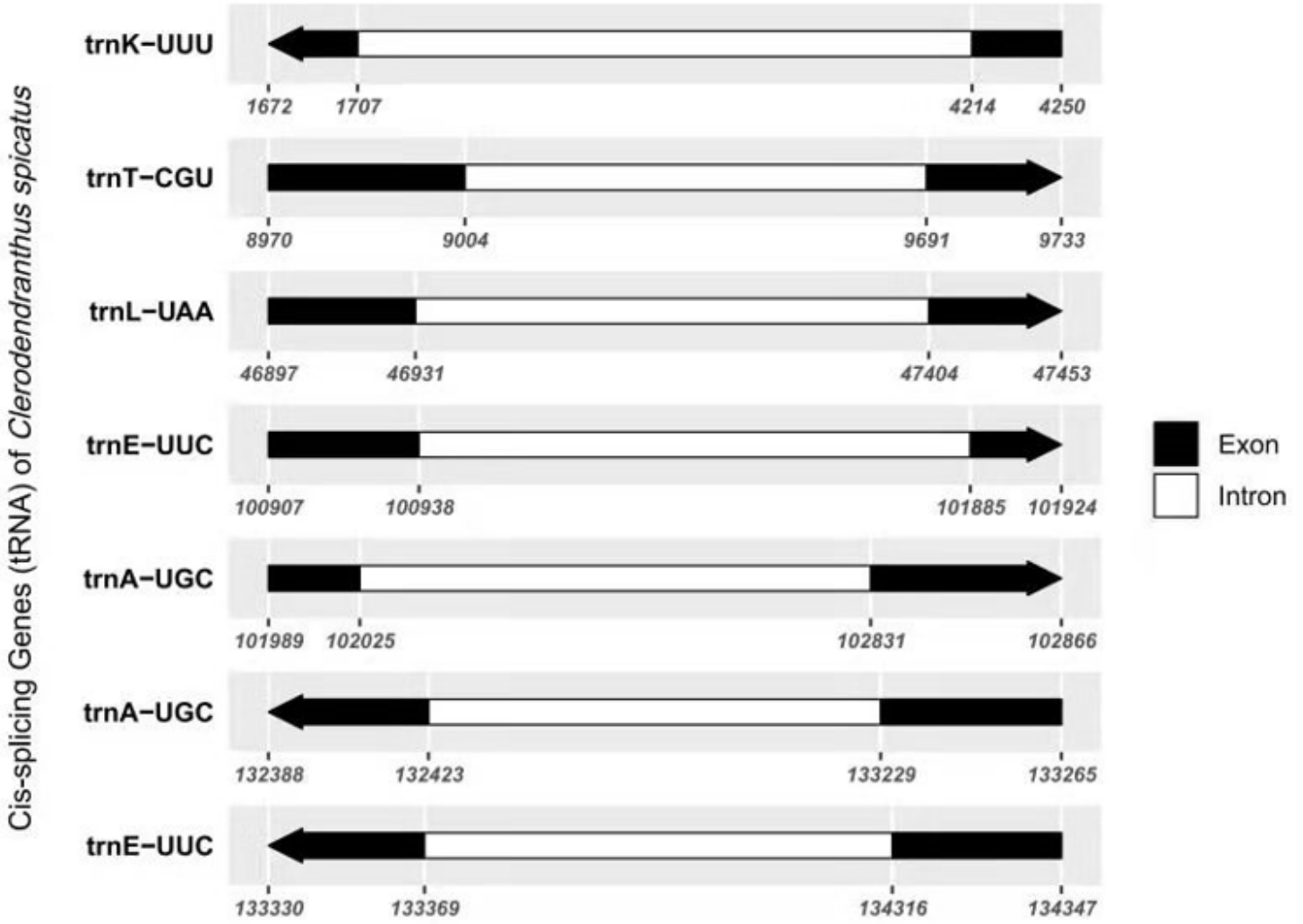

Figure 3

Schematic presentation of the structure of cis-splicing genes (tRNA) from the plastome of Clerodendranthus spicatus (http://www.herbalgenomics.org/cpgview). The white areas present introns, and the black area present exons. The arrow shows the sense direction of the genes.

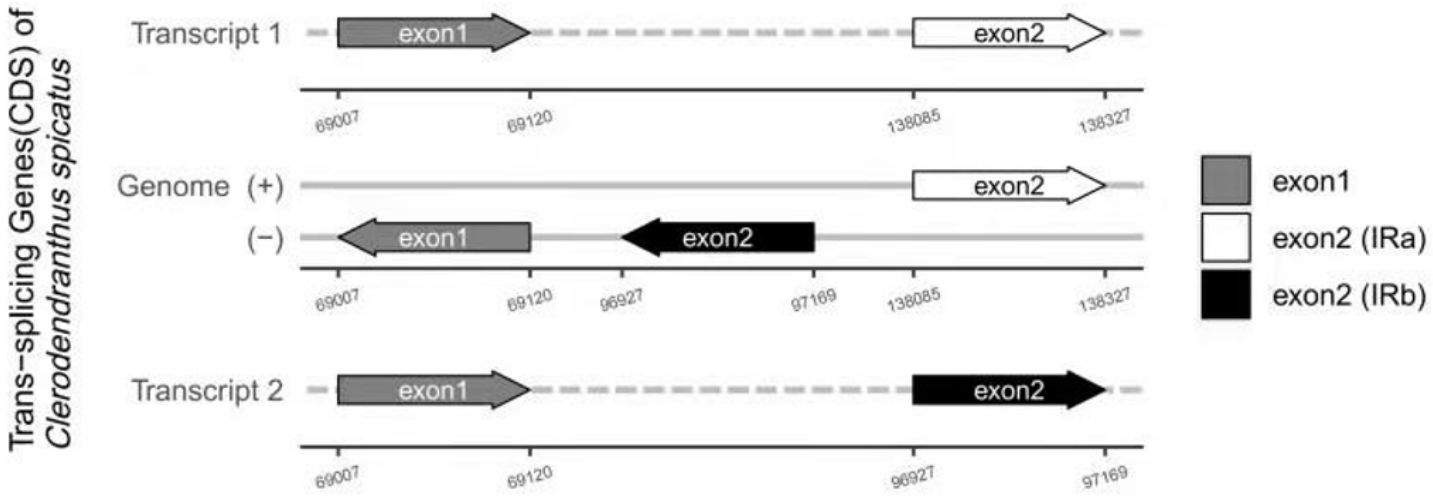

Figure 4

Schematic presentation of the structure of trans-splicing genes from the plastome of Clerodendranthus spicatus (http://www.herbalgenomics.org/cpgview). The white area is exon 2 and 3 in IRa, the black area is the exon2 and 3 in IRb and the grey area is the exon 1. The arrow shows the sense direction of the genes.

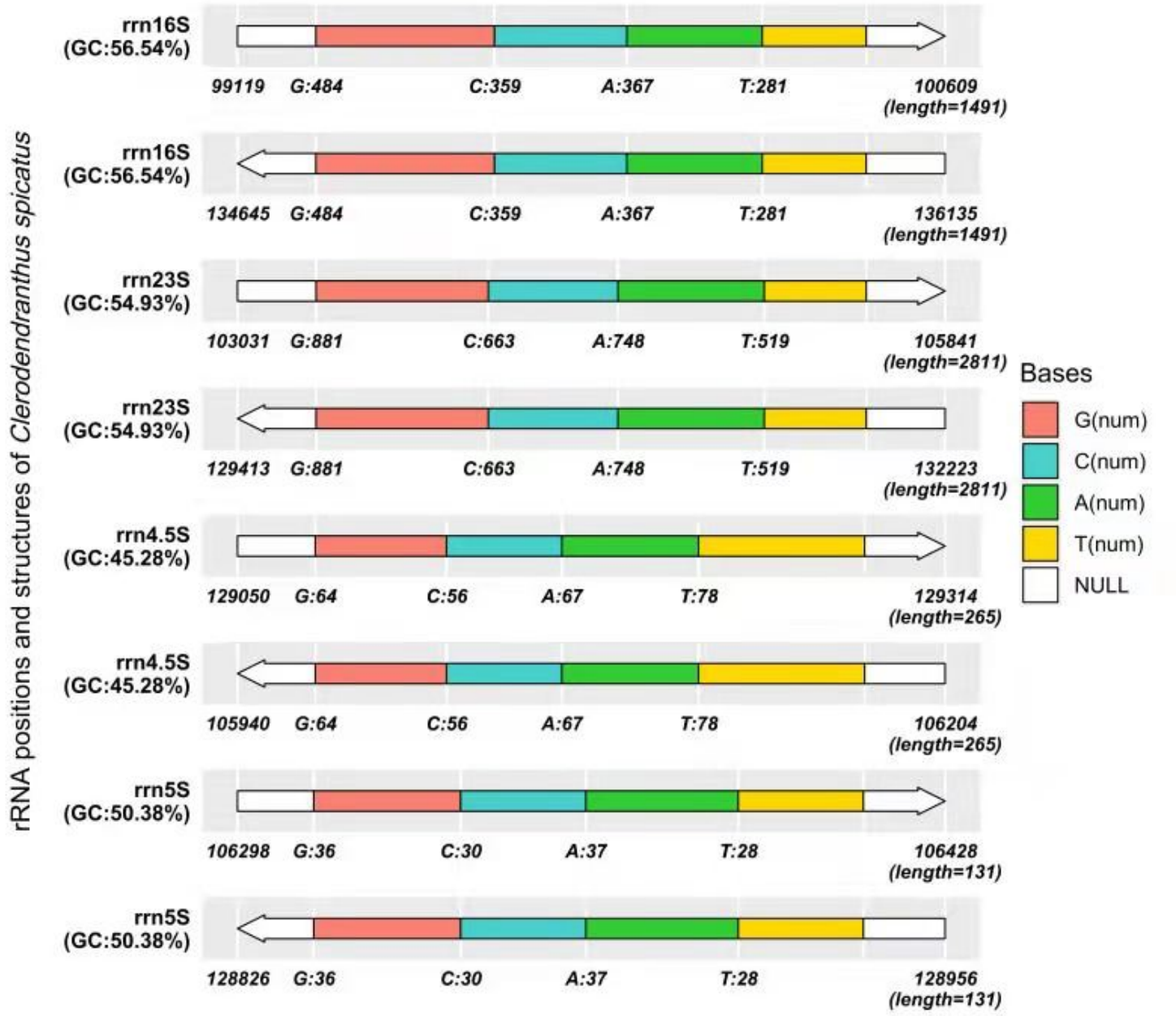

Figure 5

Eight rRNA positions and structures of Clerodendranthus spicatus chloroplast genome (http://www.herbalgenomics.org/cpgview). The numbers on the left are the start of rRNAs and GC contents of rRNAs. The numbers on the right are the end and lengths of rRNAs under the arrow line. The area of pink,blue,green and yellow are the G, C, A and T bases of diverse rRNAs. The number behind of base is the quantity of different bases. The arrow shows the sense direction of the genes.

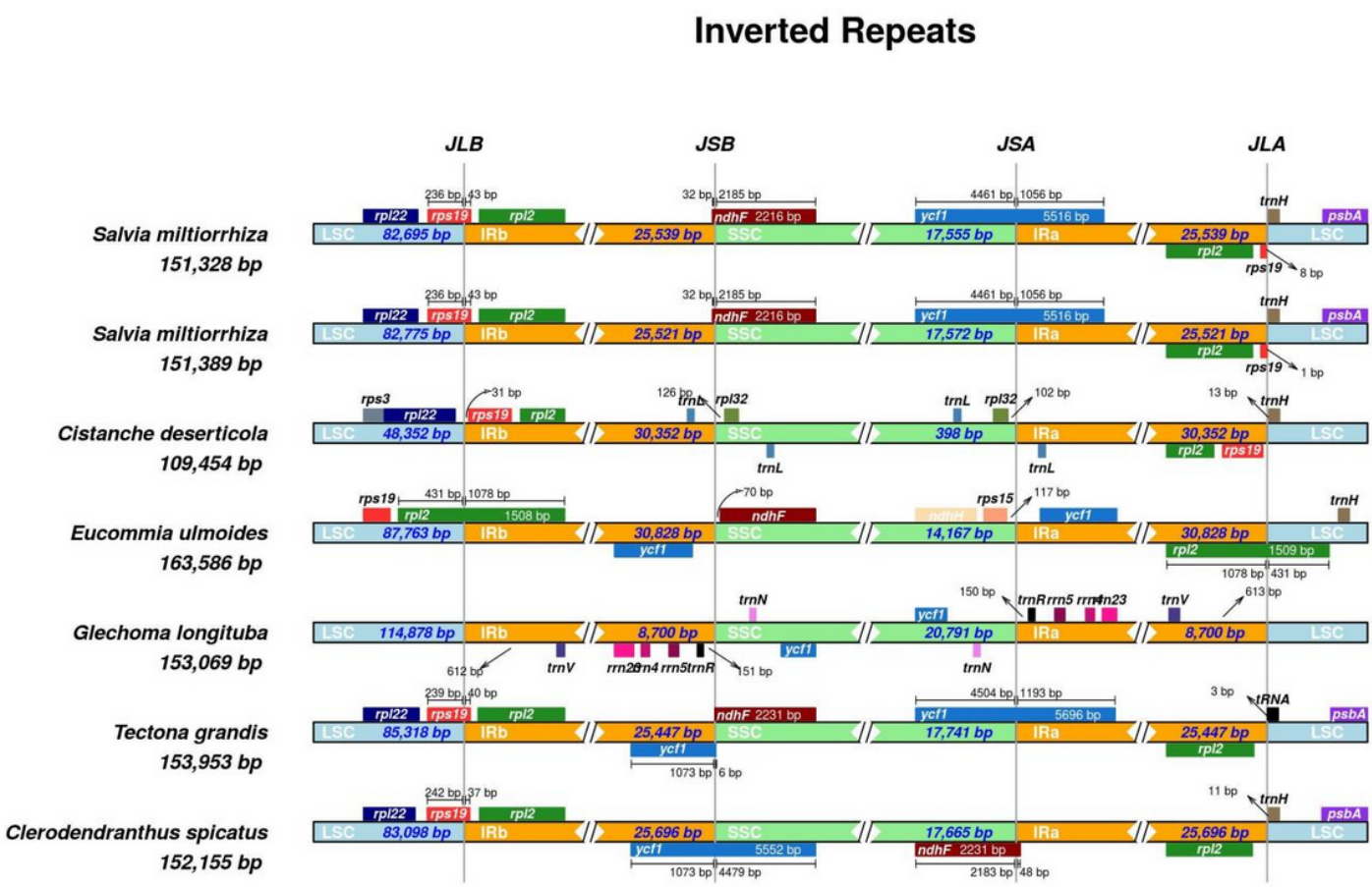

Figure 6

Comparison of the borders among large single-copy (LSC), short single-copy (SSC), and the inverted repeat (IR) regions of 7 species chloroplast genome. The genes are denoted by the colored boxes. The gaps between the genes and the boundaries are indicated by the base lengths (bp). The thin lines represent the connection points of each area, and the information of the genes near the connection points is shown by the figures. The species latin names and the length of the plastomes were shown on the left. The JLB, JSB, JSA, and JLA represent junction sites of LSC/IRb, IRb/SSC, SSC/IRa, and IRa/LSC, respectively. The distances from the start and end positions of different genes across junction sites were shown above orbelow the corresponding genes.

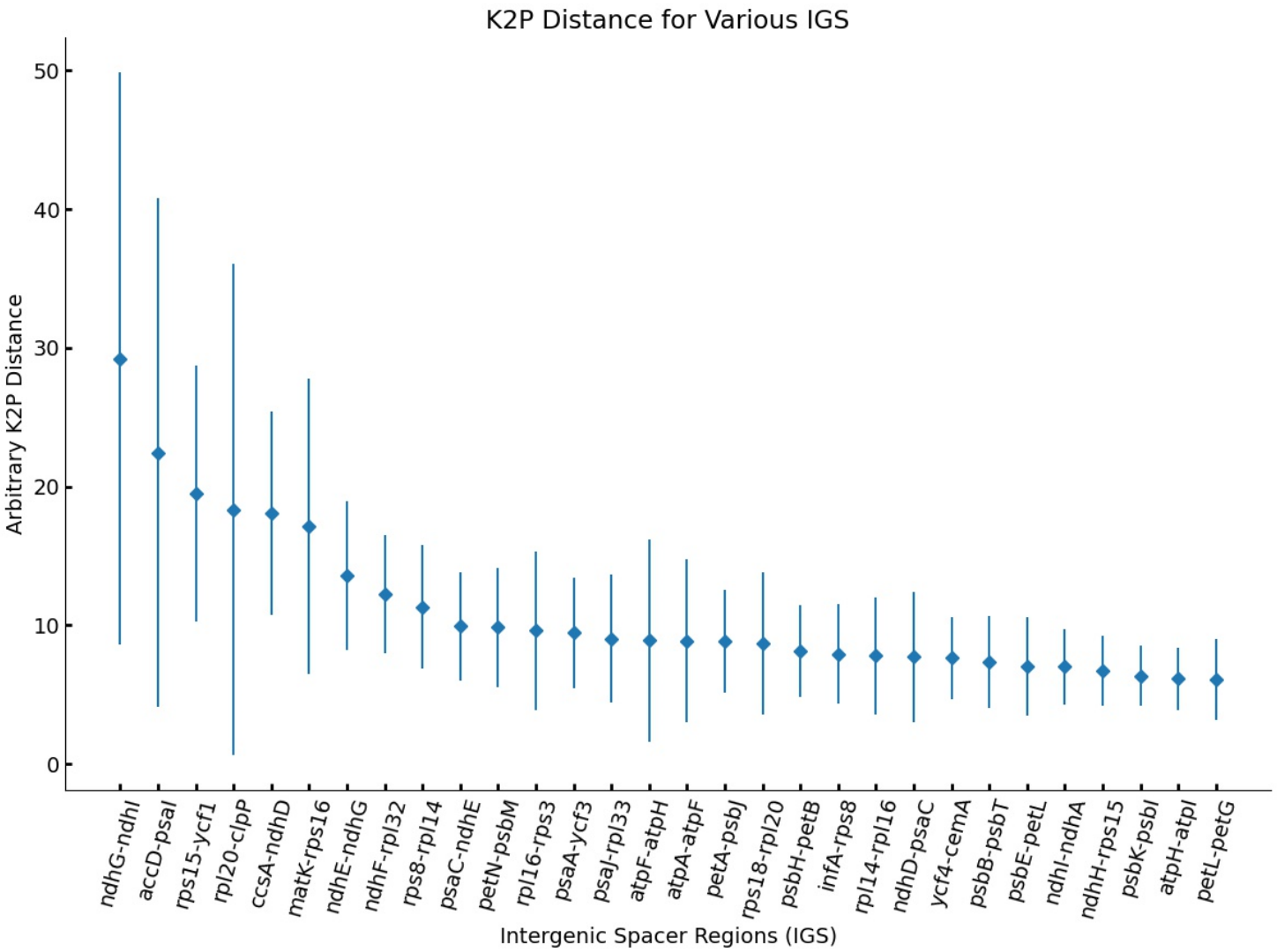

Figure 7

The phylogenetic tree of 15 (Clerodendranthus spicatus, Cistanche deserticola, Glechoma longituba, Salvia miltiorrhiza, Salvia miltiorrhiza f. alba and Tectona grandis) selected species from Lamiaceae and 1 species (Epimedium brevicornu) from Berberidaceae family as the outgroup. The tree was constructed with the sequences of 51 proteins shared in all 7 species by using the maximum likelihood method. Bootstrap supports were calculated from 1000 replicates.

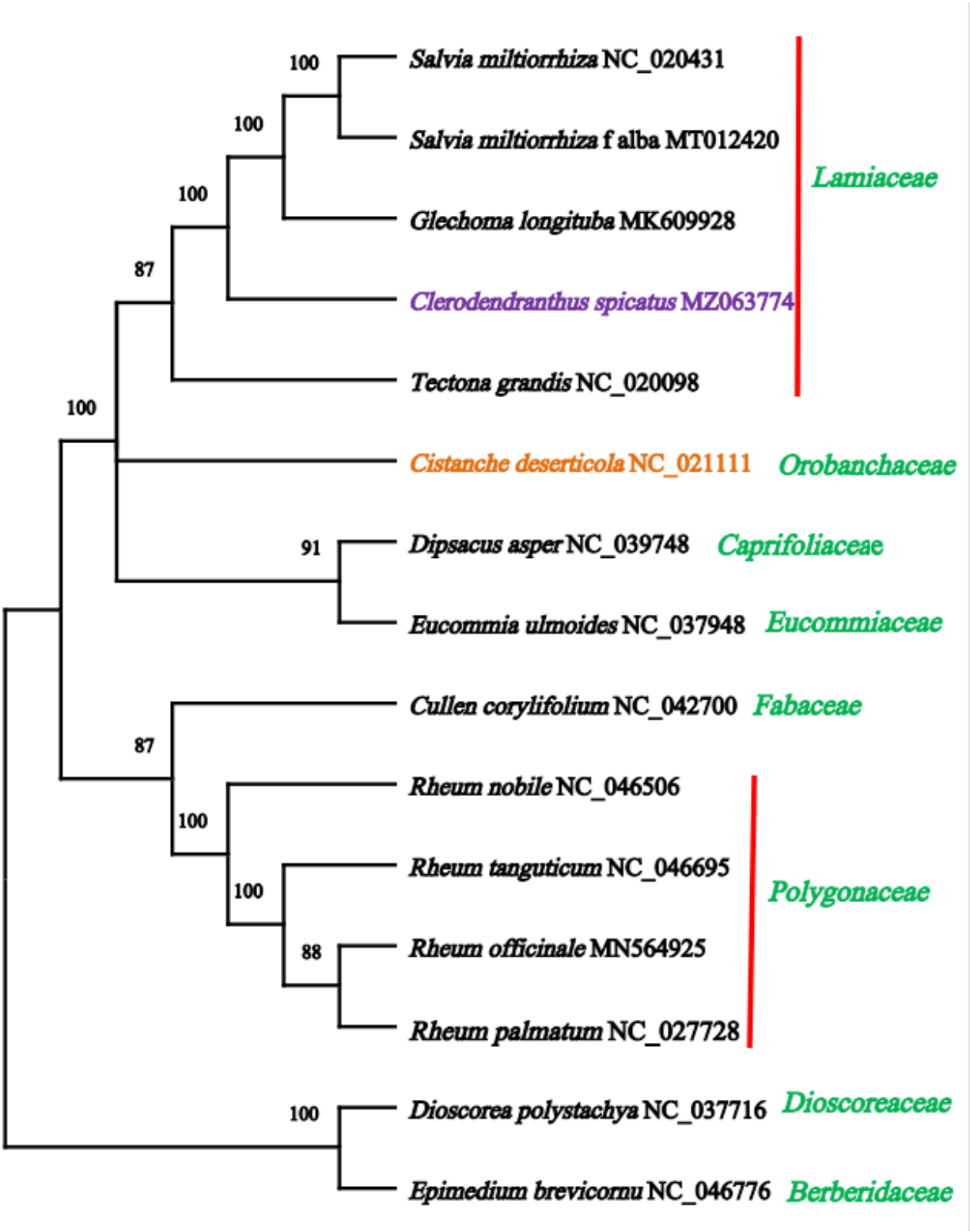

Figure 8

Average K2p distances for intergenic spacer regions among 5 selected species from Lamiaceae (Clerodendranthus spicatus, Glechoma longituba, Salvia miltiorrhiza, Salvia miltiorrhiza f. alba and Tectona grandis). The K2p distances were calculated among five chloroplast genome in pairs. The black dots represent the average value of three pairs. The Error bars represent the standard error among three pairs.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterials.docx