The complete chloroplast genome of the C. tangutica (Maxim.) Korsh. and an adaptive evolutionary analysis of the ycf2 gene

Xiaozhu Guo
Academy of Agriculture & Forestry, Qinghai University, Xining, Qinghai Province,China
Yongqiang Jiang
Academy of Agriculture & Forestry, Qinghai University, Xining, Qinghai Province,China
Shipeng Yang
Academy of Agriculture & Forestry, Qinghai University, Xining, Qinghai Province,China
Xuemei Sun
13997091543@163.com

Academy of Agriculture & Forestry, Qinghai University, Xining, Qinghai Province,China

Research Article

Keywords: Chloroplast genome, C. tangutica (Maxim.) Krosh., ycf2 gene

Posted Date: March 7th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4011223/v1

Additional Declarations: No competing interests reported.

Abstract

Clematis. tangutica (Maxim.) Korsh. is widely cultivated in Northwest China. The whole plant was called “Ye-Mang-Na-Bao” in traditional Tibetan medicine. In this study, Illumina sequencing technology was utilized to assemble and annotate the complete chloroplast genome sequences of it. The total length was 159, 584 bp, including four conserved regions: A pair of reverse repeat regions (IRa 31,042 bp and IRb 31,042 bp), a large single-copy region (79,515 bp), and a small single-copy region (17,985 bp). The genome had a total of 136 genes, with 16 presented in the reverse direction in the IR region. A total of 23 simple sequence repeats (SSRs) were identified in the coding and non-coding regions, most of which were biased toward A/T bases. A total of 15 SSRs were distributed in the non-coding regions. A comparative analysis of the chloroplast genome sequence of the *C. tangutica* (Maxim.) Krosh. and other species of the Ranunculaceae revealed that the chloroplast genome sequences of plants of the Ranunculaceae were highly conserved. Differences were observed in 13 gene loci in the coding region, with the degree of differentiation of the *ycf2* gene being the most obvious. A phylogenetic analysis showed that *Clematis glauca* had the closest relationship with *C. tangutica* (Maxim.) Krosh., both members of the Clematis genus. Selective locus detection of the *ycf2* gene in nine species of the Ranunculaceae was performed to explore adaptive evolution traits of the *ycf2* gene in it. The results show that there are significant and extremely significant positive selection sites at the 103L, 352L, 356S, 362S, 363D, 366P, 368C, 369L, 374R, 376F, 377T and 382Q loci, respectively, indicating that the *ycf2* gene has been subject to adaptive evolution.Insights from our assessment of the complete chloroplast genome sequences of *C. tangutica* (Maxim.) Krosh. will aid in the in-depth study of the evolutionary relationship of the Ranunculaceae and provide significant sequencing information for the genetic improvement of it.

Introduction

Clematis L. (Ranunculaceae)(Wen-Tsai and Liang-Qian 2005), which is one of the largest of the Ranunculceae, is a large genus that consists of 355 species in the world. A total of 155 species (93 of them endemic) are found in China, of which 70 are used as traditional Chinese medicines. China, which is believed to be the center of diversity of the genus, alone has 147 species, of which 93 are endemic(Grey-Wilson 2000). Many Clematis species are widely cultivated as ornamentals because of their beautiful corollas and unique flower shapes, and as one of the most widespread genera of flowering plants, and its ornamental value and species diversity have evoked research and exploration.

Clematis. tangutica (Maxim.) Korsh(Zhang et al. 2006; Guo et al. 2023). (Fig.1) is a perennial twining plant that is widespread in Tibet, Qinghai, Gansu and Sichuan Province. The whole plant was called “Ye-Mang-Na-Bao” in traditional Tibetan medicine and was used to treat skin ulcers, indigestion and ailments such as nervous disorders, syphilis, gout, malaria, dysentery, rheumatism, asthma, and as analgesic, anti-inflammatory, diuretic, antitumour, antibacterial and anticancer(Zhao et al. 2016).

In addition to having their own genome, chloroplasts are distinctly important organelles. By converting solar energy into carbohydrates, they sustain plant growth and development(Asaf et al. 2017). Chloroplast genomes contain valuable information and have been used as ideal research models, particularly for molecular markers, barcoding identification, plant phylogenetics, evolution, and comparative genomic studies(Wang et al. 2016). Plants convert solar energy into carbohydrates through photosynthesis to sustain growth and development(Raveendar et al. 2015). A typical circular chloroplast genome has a conserved quadripartite structure consisting of a large single-copy region (LSC) and a small single-copy region (SSC), which are separated by a pair of inverted repeats (IRs). In addition, most chloroplast genomes of angiosperms are 120–160 kb in size(Yang et al. 2010). There are two categories of chloroplast genes: protein-coding genes and non-coding genes. It is further divided into introns and intergenic

regions(Shaw et al. 2007). Chloroplast genome sequences were first reported in 1986 for tobacco (Nicotiana tabacum) and liverwort (Marchantia polymorpha)(Shinozaki et al. 1986). So far, the National Center for Biotechnology Information (NCBI) has recorded more than 960000 chloroplast genome sequences.

Materials & Methods

Samples and genome sequencing

Fresh tender leaves of *C. tangutica* (Maxim.) Korsh. were obtained from the Qinghai University Academy of Agricultural and Forestry Sciences (36°70´N, 99°54´E). Chloroplast DNA was extracted through an improved high-throughput chloroplast genome extraction method(Shi et al. 2012). Illumina HiSeq PE150 paired-end sequencing technology was used to establish the library for sequencing. The library was of the DNA small fragment type with 4.39G read length with the average depth was 972×.

Chloroplast genome assembly and annotation

FastQC was used for the quality filtering of clean data. SOAPdenovo software was used for pre-assembly(Luo et al. 2012), while SPAdes v3.14.1 was used for sequence assembly(Bankevich et al. 2012). The sequence of the chloroplast genome of *Clematis aethusifolia* (MK253462) was used as a reference to determine the location of the chloroplast genome. Extended and merge the collected fragmented target sequences using PRICE and MITObim to minimise the number of scaffolds. A reference sequence is obtained using an iterative splicing method, then the original sequencing reads are matched against this reference sequence using Bowtie2 to find matching pairs of reads, and finally these matching reads are rejoined using SPAdes. PAG software was used for annotation. The above program uses default parameters. The gene region and protein coding sequence were manually adjusted according to the initiation codon and termination codon sequences. tRNA was entered into tRNAscan-SE (http://lowelab.ucsc.edu/tRNAscan-SE/) for annotation(Lowe and Eddy 1997). rRNA was submitted to the RNAmmer 1.2 Server (http://www.cbs.dtu.dk/services/RNAmmer/) for prediction. The resulting sequence information and annotation results were submitted to Genebank, with the sequence number of OR063964. The Organellar Genome DRAW software (http://ogdraw.mpimp-golm.mpg.de/index.shtml)(Greiner et al. 2019) was used to render a complete circular chloroplast genome map.

Repeats and SSRs analysis

The dispersed repeats in chloroplast genome were detected using REPute software (https://bibiserv.cebitec.uni-bielefeld.de/reputer) with a minimum repeat length of 30 bp and a similarity threshold of 90%. Tandem repeats were detected using Tandem Repeats Finder software (https://tandem.bu.edu/trf/trf.html) with Alignment Parameters (match,mismatch,indels) of (2,7,7), Minimum Alignment Score To Report Repeat of 70, Maximum Period Size of 1000 and Maximum TR array Size (bp,millions) of 5. Simple sequence repeats (SSRs) were identified by Krait (v1.3.3) software. The minimum repeats for each perfect SSR type were set to 12 for Mono-, 7 for Di-, 5 for Tri, 4 for Tetra-, 4 for Penta-, 4 for Hexa-, and motif standardization level was set to Level 3.

Comparative analysis of different Ranunculaceae chloroplast

The LAGAN model in the mVISTA software (Dubchak and Ryaboy 2005) was used to perform a comparative analysis of the chloroplast genome of *C. tangutica* (Maxim.) Korsh. with *C. tangutica* (MK253446.1), *C. uncinata* (NC_039846.1), *C. guniuensis* (NC_050373.1). *Paeonia suffruticosa* (OK662586.1), *Paeonia mlokosewitschii* (NC_071855.1), *Trollius macropetalus* (NC_059918.1), *Trollius farreri* (NC_050872.1) and *Anemone nemorosa*

(NC_066446.1). IRscope(Amiryousefi et al. 2018) was used to indicate the boundaries between the IR and SC regions of 9 chloroplast, and the results were used as references to visualize the final view after manual revision. HomBlocks(Bi et al. 2018) was used to construct a Circos map (http://circos.ca/) to find the direction, relative position and link color of the genes. This was then standardized according to the length of all the alignment regions. Coloring was performed in accordance with the long, medium, relative short, and short sequence lengths (red, orange, green, and blue, respectively). COBALT (https://www.ncbi.nlm.nih.gov/tools/cobalt/cobalt.cgi?CMD=Web) was utilized to compare the differential protein sequence *ycf2*. HomBlocks and COBALT use default parameters.

Phylogenetic analysis

The following 25 species of the Ranunculaceae were used for the phylogenetic analysis of *C. tangutica* (Maxim.) Korsh.: *C. uncinate* (NC_039846.1), *C. terniflora* (KJ956785.1), *C. guniuensis* (NC_050373.1), *C. heracleifolia* (NC_039845.1), *C. tangutica*(MK253446.1), *C. intricate* (NC_065278.1), *C. glauca* (ON520701.1), *C. florida* (NC_058885.1), *Paeonia suffruticosa* (OK662586.1), *Paeonia lactiflora* (MK860971.1), *Paeonia mlokosewitschii* (NC_071855.1), *Paeonia sinjiangensis* (NC_071854.1), *Trollius macropetalus* (NC_059918.1), *Trollius chinensis* (NC_031849.1), *Trollius farreri* (NC_050872.1), *Anemone hortensis* (NC_066445.1), *Anemone nemorosa* (NC_066446.1), *Anemone shikokiana* (NC_069812.1), *Arabidopsis thaliana* (NC_000932.1), *Solanum lycopersicum* (NC_007898.3), *Bidens conjuncta* (NC_047263.1), *Bidens campylotheca* (NC_047270.1), *Bidens valida* (NC_047256.1), *Bidens cervicata* (NC_047264.1) and *Bidens wiebkei* (NC_047262.1). Mafft was used to compare 26 chloroplast genome sequences. A phylogenetic tree was constructed with the methods of maximum-likelihood. The automatic selection of the best model was set up in the IQ Tree, and IQ Tree v8.1.24(Minh et al. 2020) was used to construct the tree. Parameters were set to search for 30 repeats, and the tree with the maximum likelihood value was used. In addition, Bootstrap was set to run 1,000 times to calculate the support of each branch. To build the Bayesian tree, the nucleotide substitution model GTR+I+G in Bayesian analysis was selected according to BIC in the jModelTest 2.1.7 software(Darriba et al. 2012). MrBayes 3.2(Huelsenbeck and Ronquist 2001) was used for calculations, employing the Markov chain Monte Carlo methodology. Four Markov chains were initialized at the same time. The random tree was marked as the initial tree, and one was saved every 500 trees for a total of 5,000,000 trees. The first 20% of the burn-in trees were discarded. The remaining trees were used to calculate the posterior probability of the consistent tree and each branch.

Adaptive evolution traits

The ratio (ω) of the non-synonymous substitution (dN) to the synonymous substitution (dS) of nucleotides is used in most adaptive evolution studies to measure the selection pressure at the nucleic acid or protein level. In addition, the selection pressure is considered to hinder or promote its role in the process of non-synonymous replacement fixation. The positive selection model (M2a, M8) and the control model (M1a, M7, M8a) provided by EasyCodeML software were used to conduct the adaptive evolution analysis in the loci(Gao et al. 2019). The locus model was used to assume that there were different selection pressures at different loci. In other words, the ω values were different, but there was no difference in the different branches of the phylogenetic tree. This model was primarily used to detect the existence of positive selection (ω > 1) and negative selection (ω < 1) loci in the *ycf2* gene. Three pairs of comparison models were M1a (near neutral) and M2a (selection), M0 (single ratio) and M3 (discrete), M7 (beta) and M8 (beta & ω) in this study. The former is a zero hypothesis, and the latter is an alternative hypothesis. Models M0 (single ratio) to M3 (discrete) were used to detect different ω values at each point rather than detecting positive selection loci. PAMLx V1.3.1 was used to perform the likelihood ratio test (LRT) in three pairs of models(Xu and Yang 2013). Positive selection loci were tested by comparing the significance of the differences between the models. $\chi 2$

distribution was used as the significance test under the condition of relative degrees of freedom (the difference between the number of two models).

Results

Genome organization and gene features

The chloroplast genome of *C. tangutica* (Maxim.) Korsh. had a total length of 159,584 bp. The genome was composed of four parts: A pair of reverse repeat regions, IRa (31,042 bp) and IRb (31,042 bp), separated by a large single-copy region LSC (79,515 bp) and a small single-copy region SSC (17,985 bp) (Fig. 2) Genes in the coding regions accounted for 48.51% of the genome, including protein-coding genes, tRNA genes, and rRNA genes. The chloroplast genome of *C. tangutica* (Maxim.) Korsh. had a total guanine-cytosine content (GC content) of 37.9%, with GC in the IR region corresponding to 42.0%, and GC in the LSC and SSC regions being 36.3% and 31.4%, respectively. The chloroplast genome of *C. tangutica* (Maxim.) Korsh. contained 136 genes, including 92 protein-coding genes CDS, 36 tRNA genes and 8 rRNA genes distributed in the IR region. Furthermore, this region encompassed 16 inverse genes, including eight CDS genes (*ycf2*, *ndhB*, *rps2*, *rps3*, *rps12*, *rpl2*, *rpl22*, and *rpl23*), 6 tRNA genes, and 2 rRNA genes. The 136 genes contained 72 Protein synthesis and DNA replication genes, 49 Photosynthesis genes, 8 Miscellaneous group genes and 7 pseudogenes of unknown function (Table 1).

In the chloroplast genome of *C. tangutica* (Maxim.) Korsh., 18 intron-containing genes were annotated, 11 of which were protein-encoding and 6 were tRNA genes. Of the 17 intron genes, the intron sequence in *trnK-UUU* was the longest (2,545 bp), while the intron in the *petB* gene was the smallest (483 bp). There were two introns in the *clpP* and *ycf3* genes, whereas the other genes contained only one intron (Table 2). The *rps12* gene has a trans splicing condition.

Table 1
List of genes in the choroplast genome of *C. tangutica* (Maxim.) Korsh.

	Groups of genes	Names of genes
Protein synthesis and DNA replication	Ribosomal RNAs	*rrn4.5(2×)*, *rrn5(2×)*, *rrn23(2×)*, *rrn16(2×)*
	Transfer RNAs	*trnS-GGA*, *trnF-GAA*, *trnW-CCA*, *trnV-GAC(2×)*, *trnC-GCA*, *trnR-ACG(2×)*, *trnG-GCC*, *trnL-CAA(2×)*, *trnQ-UUG*, *trnL-UAG*, *trnD-GUC*, *trnS-GCU*, *trnE-UUC*, *trnS-UGA*, *trnY-GUA*, *trnV-UAC*, *trnP-UGG*, *trnH-GUG*, *trnN-GUU(2×)*, *trnK-UUU*, *trnL-UAA*, *trnfM-CAU*, *trnI-CAU(2×)*, *trnT-GGU*, *trnR-UCU*, *trnG-UCC*, *trnV-UAC*, *trnK-UUU*, *trnL-UAA*, *trnI-GAU(2×)*, *trnG-UCC*, *trnM-CAU*, *trnI-GAU(2×)*, *trnA-UGC(2×)*, *trnA-UGC(2×)*
	Ribosomal protein small subunit	*rps7(2×)*, *rps14*, *rps12(2×)*, *rps2*, *rps4*, *rps12(2×)*, *rps11*, *rps16*, *rps12(2×)*, *rps3(2×)*, *rps15*, *rps8(2×)*, *rps7*, *rps4*, *rps16*, *rps19(2×)*, *rps18*
	Ribosomal protein large subunit	*rpl14(2×)*, *rpl23(2×)*, *rpl36*, *rpl2(2×)*, *rpl20*, *rpl2(2×)*, *rpl32*, *rpl16(2×)*, *rpl33*, *rpl22(2×)*, *rpl16(2×)*
	Subunits of RNA polymerase	*rpoB*, *rpoA*, *rpoC2*, *rpoC1*, *rpoC1*
Photosynthesis	Photosystem I	*psaC*, *psaA*, *psaB*, *psaI*, *psaJ*
	Photosystem II	*psbZ*, *psbK*, *psbB*, *psbI*, *psbF*, *psbN*, *psbL*, *psbJ*, *psbC*, *psbE*, *psbM*, *psbH*, *psbA*, *psbD*, *psbT*
	Cytochrome b/f complex	*petA*, *petD*, *petL*, *petB*, *petG*, *petN*, *petD*, *petB*
	ATP synthase	*atpE*, *atpH*, *atpA*, *atpI*, *atpF*, *atpB*, *atpF*
	NADH-dehydrogenase	*ndhA*, *ndhJ*, *ndhG*, *ndhI*, *ndhH*, *ndhE*, *ndhD*, *ndhC*, *ndhF*, *ndhA*, *ndhB(2×)*, *ndhB(2×)*,*ndhK*
	Large subunit Rubisco	*rbcL*
Miscellaneous group	Translation initiation factor IF-1	*infA*
	Acetyl-CoA carboxylase	*accD*
	Cytochrome c biogenesis	*ccsA*
	Maturase	*matK*
	ATP-dependent protease	*clpP*, *clpP*, *clpP*,
	Inner membrane protein	*cemA*

	Groups of genes	Names of genes
Pseudogenes of unknown function	Conserved hypothetical chloroplast open reading frame	*ycf4*, *ycf3*, *ycf3*, *ycf3*, *ycf1*, *ycf1*, *ycf2(2×)*

Table 2
Characteristics of genes including introns and exons in the chloroplast genome of *C. tangutica* (Maxim.) Korsh.

Gene	Region	Exon I (bp)	Intron I (bp)	Exon II (bp)	Intron II (bp)	Exon III (bp)
rps	LSC	40	882	221		
trnK-UUU	LSC	37	2545	35		
trnL-UAA	LSC	35	491	50		
trnG-UCC	LSC	23	717	48		
atpF	LSC	410	740	145		
rpoC1	LSC	1613	768	430		
trnV-UAC	LSC	35	583	38		
petB	LSC	6	483	642		
petD	LSC	8	731	496		
rpl16	IR	399	977	9		
rpl2	IR-IR	434	651	391		
ndhB	IR	756	708	777		
trnI-GAU	IR	37	947	35		
trnA-UGC	IR	38	813	35		
ndhA	SSC	539	1001	553		
clpP	LSC	246	699	289	795	71
ycf3	LSC	152	743	230	726	124

Repeats and SSRs analysis

A total of 26 dispersed repeats were detected, which met the criteria of a hamming distance of 3, a maximum computed repeats of 5000 and a minimal repeat size of 30. These included 8 reverse complement (P), 14 forward repeats (F), 1 complement repeat (C), and 3 reverse repeats (R). Ten tandem repeats in the length range of 9–24 bp were also detected, distributed over four regions, with two and three repeats in the LSC and SSC regions, respectively, and three in each of the two IR regions, with a repeat count of two. The distribution of chloroplast simple sequence repeat (cpSSR) in *C. tangutica* (Maxim.) Korsh. was analyzed, revealing 23 different SSR loci in its chloroplast genome, and 8 different SSR loci in the coding region which are found in the *ycf1*, *psbC*, *rpoA*, *rpoB* and *rpoC2*.

Among them, 21 SSR were composed of A, 1 were composed of AT, and only one was composed of AAC, indicating that the chloroplast genomic SSR of *C. tangutica* (Maxim.) Korsh. are biased toward A bases (Fig. 3, Table.3). An assessment of the SSR distribution identified 15 SSR in the non-coding region of the chloroplast genome. In the coding region, SSR are only found in the *ycf1*, *psbC*, *rpoA*, *rpoB* and *rpoC2* genes.

Table 3
The abundance of each standard motif category

Motif	Type	Counts	Length(bp)	Percent (%)	Average length (bp)	Relative Abundance (loci/Mb)	Relative Density (bp/Mb)
A	1	21	270	91.3	12.86	131.59	1691.9
AT	2	1	26	4.35	26.0	6.27	162.92
AAC	3	1	18	4.35	18.0	6.27	112.79

Comparative analysis of different composite chloroplast

A comparative analysis with the plastomes of other species (eespecially with the same) of the Ranunculaceae revealed only small differences in chloroplast size and composition in comparison to that of *C. tangutica* (Maxim.) Korsh. (Table 4). There were very few inconsistencies in the types and number of chloroplast genes in several species of the Ranunculaceae, and the types and number were very conserved. The total size chloroplast genome of *C. tangutica* (Maxim.) Korsh. ranked 5th in the aligned genomes of the 9 chloroplast genomes of the genus Clematis. The variation in the length of the sequence may be caused by the difference in length between the LSC and IR regions. The chloroplast genome size of 9 crops of the Ranunculaceae was approximately 160 kb, with a GC content of approximately 38%. The number of protein-coding genes ranged between 83 and 92. All of these genomes had 8 rRNA-coding genes and 36–39 tRNA-coding genes. The plastome of *C. tangutica* (Maxim.) Korsh. was 135 bp longer than that of *C. tangutica* (a crop in the same genus), primarily in the LSC region. In addition, it had one more protein-coding genes than that of *C. tangutica*, with no difference in the number of rRNA- and tRNA-coding genes.

Table 4
Comparison of cp genomes among eight composite species.

Species	Siza(bp)				G + C (%)	Total number of genes			GeneBank accessions
	Total	LSC	IR	SSC		Protein-coding genes	rRNAs	tRNAs	
C. tangutica (Maxim.) Korsh.	159584	79515	62084	17985	37.90	92	8	36	
C.tangutica	159449	79291	62090	18118	37.95	91	8	36	MK253446.1
C.uncinata	159524	79347	62078	18099	38.02	90	8	36	NC_039846.1
C. guniuensis	159682	79461	62128	18093	37.95	86	8	36	NC_050373.1
P. suffruticosa	152596	84272	51280	17044	38.37	85	8	37	OK662586.1
P. mlokosewitschii	152785	84450	51310	17025	38.43	83	8	37	NC_071855.1
T. macropetalus	160094	88555	53248	18291	38.03	87	8	39	NC_059918.1
T. farreri	160611	88943	53136	18532	38.05	87	8	36	NC_050872.1
A. nemorosa	160895	80853	62396	17646	37.66	90	8	36	NC_066446.1

The genomic sequences of 9 composite species were analyzed by the mVISTA software, detecting the variations of the sequences (Fig. 4). The results showed there was less variation between *C. tangutica* (Maxim.) Korsh., *C. tangutica* and *C. uncinata* and *C. guniuensis*. Compared with *Clematis florida* Thunb., several structures were lacking in the *C. tangutica* (Maxim.) Korsh..

Compared the IR boundaries of nine species, based on phylogenetic analyses. Figure 5 shows that the majority of them share a similar structure in the same genus. The *ycf1* gene are found embedded in IR region and SSC region. In the *C. tangutica* (Maxim.) Korsh., *C. tangutica*, *C. uncinate*, *C. guniuensis* and *A. nemorosa*, the rps11 is completely covered by LSC region and rps8 is compelely covered by IRb region. The rps4 gene are found fully embedded in LSC region for *C. tangutica* (Maxim.) Korsh., *C. tangutica*, *C. uncinate* and *A. nemorosa*. The *rps8* gene are found fully embedded in IRa region for *C. tangutica*, *C. uncinate* and *C. guniuensis*. The *infA* gene was found embeddedin IRb region and expanded to LSC region for *C. tangutica* (Maxim.) Korsh., *C. tangutica* and *C. uncinate*. The IRb/SSC boundary is the most conserved, while all the *ycf1* gene of *C. tangutica* (Maxim.) Korsh., *C. uncinate*, *P. suffruticosa*, *T. farreri* and *A. nemorosa* are located in the SSC region and IRb, however the location of *ycf1* gene in the SSC region varies in these species, ranging from 2 to 49 bp.

Based on the results of mVISTA, a systematic comparative analysis was performed in a coding region with small variation amplitude(Doorduin et al. 2011). As shown in Fig. 6, there were differences among nine species of the Ranunculaceae in the following 13 gene loci: *rps7*, *ndhB*, *ycf2*, *rpl23*, *rpl2*, *rps19*, *rpl22*, *rps3*, *rpl16*, *rpl14*, *rps8*, *infA*, and *rps12*. The discovery of these differential genes provides valuable phylogenetic information for the further evaluation of the Ranunculaceae.

In many studies, the *ycf2* gene has become an alternative choice for the assessment of plant sequence variation and phylogenetic evolution. Our results showed that the *ycf2* gene segment had a large deletion and inconsistency. The *ycf2* gene of *C. tangutica* (Maxim.) Korsh. and eight other Ranunculaceae was compared. Four species of the genus

Clematis had 26 amino acid sequence deletions of the *ycf2* gene in the segment 360–383 (Fig. 7). There were 94 amino acid sequence deletions in segment 400–500 of *P. mlokosewitschii* and *P. suffruticosa*. In addition, there were some amino acid site differences. Lastly, the greatest similarity was observed between the *ycf2* genes of *C. tangutica* (Maxim.) Krosh. and *C. guniuensis*, *C. uncinate*, *C. tangutica*.

Phylogenetic analysis

To assess the phylogenetic relationships of *C. tangutica* (Maxim.) Korsh., the chloroplast genomes of 26 species of the Ranunculaceae were compared globally. Five species of Bidens genus, *S. lycopersicum*, four species of Paeonia genus and *A. thaliana* were used as outgroups, and then Bayesian evolutionary trees were constructed (Fig. 8). All the species in the Ranunculaceae formed four highly supported evolutionary clades: members of the genus Clematis, Anemone and Trollius are included in the first clade, the genus Clematis including *C. uncinate*, *C. terniflora*, *C. florida*, *C. guniuensis*, *C. heracleifolia*, *C. tangutica* (Maxim.) Krosh., *C. glauca*, *C. intricate*, *C. tangutica*. On the evolutionary subclade of the genus Clematis, *C. tangutica* (Maxim.) Korsh. and *C. glauca* are in the closest relationship. The common node bootstrap is fully resolved. Bidens and Solanum are contained in the second clade, while genus Paeonia are clustered in the third clade and *A. thaliana* is clustered alone in the Arabidopsis.

Estimation of the positive selection loci of the ycf2 gene in nine species of the Ranunculaceae

EasyCodeML v1.2 were used to calculate the logarithmic likelihood value (lnL) and parameter evaluation for the complete sequence data set of the *ycf2* coding region of nine species in the Ranunculaceae. In the locus model, $\omega > 1$ was allowed in the models M3 (discrete), M2a (selection), and M8 (beta & ω) to assume that the corresponding zero hypothetical models were the M1a (near neutral) model, M0 (one-ratio) model and M7 (beta) model. The M3, M2a and M8 models were significantly superior to their corresponding hypothetical models M0, M1a, M7 and M8a ($p < 0.01$), indicating that there were differences in the selection pressure among the points. After LRT testing, it was found that both M7 vs. M8 and M8a vs. M8 were more consistent with the analyzed data than their original hypothetical models (Table 5), and their original hypothetical models were rejected at a significant level of $p = 0.01$. A consistent positive selection locus, Positive selection sites for models M2a and M8 were 103L, 352L, 356S, 362S, 363D, 366P, 368C, 369L, 374R, 376F, 377T, 382Q, 686D, 1095L and 2120H at the 95% and 99% levels respectively by Naïve empirical bayes (NEB) (STable 6). There was 15 positive selection locus 103L, 352L, 356S, 362S, 363D, 366P, 368C, 369L, 374R, 376F, 377T, 382Q, 686D, 1095L and 2120H in the M2a model and 13 positive selection loci 103L, 352L, 356S, 362S, 363D, 366P, 368C, 369L, 372D, 374R, 376F, 377T, and 382Q in the M8 model according to a Bayes Empirical Bayes analysis. Overall, the posterior probabilities of 103L, 352L, 356S, 362S, 363D, 366P, 368C, 369L, 374R, 376F, 377T and 382Q in the NEB analysis of the M2a and M8 models were greater than 95%. Currently, this type of gene has substantial potential for application and diverse functions in the field of plant phylogeny according to the research progress of the chloroplast *ycf* gene family.

Table 5
Likelihood ratio statistics of positive selection models against their null models (2Δ ln L).

Comparison between models	2ΔlnL	d.*f*.	*p*-value
M0 vs. M3	136.3408	4	0.0000 < 0.01
M1a vs. M2a	83.3817	2	0.0000 < 0.01
M7 vs. M8	88.7187	2	0.0000 < 0.01
M8a vs. M8	83.2980	1	0.0000 < 0.01

Discussion

This may be due to the presence of four rRNA genes with high GC content in the Jerusalem artichoke IR region(Asaf et al. 2016). Conservation in the IR regions was higher than in the LSC and SSC regions due to the high G-C content(Yang et al. 2014). *C. tangutica* (Maxim.) Krosh. had chloroplast genes that were similar to those in other crops of the genus Clematis. The results of this study will be helpful for future studies on chloroplast genome evolution and positive selection research. Our analysis of the chloroplast genome of *C. tangutica* (Maxim.) Korsh. was based on these sequencing results.

During selective gene splicing, introns play an important role. The chloroplast genome was simple, relatively conserved, and maternal, which made chloroplast SSRs highly efficient molecular markers. Crossbreeding, biogeography, and population genetics studies have widely used cpSSRs. It is consistent with the chloroplast genomes of the majority of angiosperms. As for repeat length, most SSRs were 12–26 bp, indicating that the SSR segment of *C. tangutica* (Maxim.) Krosh. chloroplast genome is short. This repeated sequence might increase genetic diversity in a population by promoting the rearrangement of chloroplast genomes. The SSR sites distributed in the non-coding region are the majority, while only eight genes in the coding region have SSR sites, and there are few SSR sites in the coding region of the chloroplast genome, as has been confirmed in *Quercus* and *Saxifragaceae*. The repetitive structures of *C. tangutica* (Maxim.) Krosh. provide valuable information for the future development of molecular markers for studying phylogenetic evolution and population genetics.

C. tangutica (Maxim.) Krosh. and *A. nemorosa* showed the least amount of variation in the chloroplast genome of Ranunculaceae plants. The chloroplast genome of crops in the Ranunculaceae is generally conserved. An mVISTA analysis showed that the coding region was more conserved than the non-coding region, which is consistent with reports on crops in the Ranunculaceae, such as *Cynara cardunculus* and *Ageratina Adenophora*. *Ycf2* showed the greatest degree of differentiation. At present, many different gene regions are considered potential tools for phylogenetic analysis. In this species, DNA domains will be critical to molecular phylogeny. *Ycf2* is the largest chloroplast gene known to date. *Ycf2* can be used to predict phylogenetic relationships, but its function is unknown. In the Ranunculaceae, the *ycf2* gene is highly conserved throughout speciation. It appears that the *ycf2* gene gradually degenerates in rice and wheat, with only 734 bp remaining. A phylogenetic tree analysis of the partial angiosperm *ycf2* genes was consistent with a phylogenetic tree analysis of the whole chloroplast genome data. The evolutionary evaluation is now even more precise.

Ranunculaceae is a large angiospermic family, distributed worldwide especially in temperate and subtropical regions, and the chloroplast genome plays an important role in plant classification and phylogenetic analysis. Research on the phylogeny of Ranunculaceae crops has been plentiful. Angiosperms of the Ranunculaceae are often considered the most primitive. In the group of the Ranunculaceae in which the number of involved species is more than or equal to 2, it can be seen that genetically *C. tangutica* (Maxim.) Krosh. is more closely related to other species of Ranunculaceae, such as genus Anemone. As a result, a theoretical basis is provided for the further study of the phylogenetic relationships between phylogenetic branches of *C. tangutica* (Maxim.) Krosh. in the Ranunculaceae.

The *ycf2* gene fragment is large, and the function of its open reading frame fragment is not clear. Compared with other chloroplast coding genes, the nucleotide sequence identity between *ycf2* of different families is very low, which is less than 50% in bryophytes, pteridophytes, and spermatophytes(Wicke et al. 2011). In the increasing number of *ycf* gene studies, although *ycf2* is highly conserved, the *ycf2* gene shows a wealth of phylogenetic information in the Orchidaceae phylogeny. One study(Huang et al. 2010) found that the *ycf2* gene has multiple positive selection loci during angiosperm development, and the phylogenetic signal of *ycf2* probably originates from its large sequence

length, so that the *ycf2* gene is valuable for future research. Most chloroplast genes were in a negative selection state in Holcoglossum, but 14 positive selection loci were detected in the *ycf2* gene(Li et al. 2019). In this study, some positive selection signals were found by establishing evolutionary trees of the adaptive evolution of the *ycf2* gene in the Ranunculaceae. Plants may have a variety of strategies to adapt to the environment, and adaptive modifications to other abiotic stresses of genes in the nucleus are sufficient to maintain the homeostasis of photosynthesis. Therefore, there is no need for adaptive evolution in the chloroplast coding genes(Kanzaki et al. 2017). In this study, research on the *ycf2* gene in the Ranunculaceae supports the idea of adaptive evolution, but there are currently few studies on adaptive evolution in Ranunculaceae. Therefore, further studies on the adaptive evolution of chloroplast genes in other species of the Ranunculaceae are needed to explore how to adapt to these changes in environmental migration and climate change.

Conclusions

C. tangutica (Maxim.) Krosh. chloroplast genome sequence was successfully assembled, annotated, and analyzed in this study. The chloroplast genome of the plants in the Ranunculaceae is relatively conserved. Variations of the chloroplast genome are scarce between *C. tangutica* (Maxim.) Krosh. and plants in the same genus. Compared with Ranunculaceae belonging to other genera, we found deletions in the chloroplast genome of *C. tangutica* (Maxim.) Krosh.. The identification of repetitive sequences in the chloroplast genome of it, particularly SSR, will be helpful for the development of molecular markers, the study of population genetics and the phylogenetic analysis of it. A phylogenetic analysis of plants in the Ranunculaceae shows that *C. tangutica* (Maxim.) Krosh. and *C. glauca* share the closest relationship, both belonging to the Ranunculaceae, genus Clematis. The results of this study indicate *ycf2* gene has been subject to adaptive evolution, and it is suggested that more extensive investigation and in-depth discussion should be conducted in future studies. Completion of the sequencing of the chloroplast genome will provide key genetic information for further research on *C. tangutica* (Maxim.) Krosh. and deepen our understanding on the evolutionary history of the chloroplast genome and phylogenetic position of it. In addition, it may be useful for various molecular biology applications of *C. tangutica* (Maxim.) Krosh. in the future.

Declarations

We would like to thank anonymous reviewers for improving our manuscript through their insightful comments and suggestions.

Additional information

Competing interests

The authors declare no competing interests.

Ethical statement

No ethical statement was reported.

Funding

This work was supported by the Qinghai Innovation Platform Construction Program (2020-ZJ-Y02).

Authors and Affiliations

Academy of Agriculture & Forestry, Qinghai University, Xining, Qinghai Province,China

Xiaozhu Guo, Yongqiang Jiang, Shipeng Yang, Xuemei Sun

Laboratory for Research and Utilization of Germplasm Resources in Qinghai Tibet Plateau, Xining, Qinghai Province, China

Xiaozhu Guo, Yongqiang Jiang, Shipeng Yang, Xuemei Sun

Contributions

Xiaozhu Guo: Methodology, Investigation, Data processing, Writing-Original draft preparation and manuscript finalization.

Yongqiang Jiang: Methodology, Investigation, Data processing, Writing-Original draft preparation and manuscript finalization.

Shipeng Yang: Methodology, Data processing, Writing-Reviewing and Editing.

Xuemei Sun: Conceptualization, Resources, Funding acquisition, Supervision, Writing-Reviewing and Editing and manuscript finalization.

Corresponding authors

Correspondence to Xuemei Sun.

References

1. Amiryousefi A, Hyvönen J, Poczai P (2018) IRscope: an online program to visualize the junction sites of chloroplast genomes. Bioinformatics 34: 3030–3031. https://doi.org/10.1093/bioinformatics/bty220.
2. Asaf S, Khan AL, Khan AR, Waqas M, Kang S-M, Khan MA, Lee S-M, Lee I-J (2016) Complete chloroplast genome of Nicotiana otophora and its comparison with related species. Front Plant Sci 7: 843. https://doi.org/10.3389/fpls.2016.00843.
3. Asaf S, Khan AL, Khan MA, Waqas M, Kang S-M, Yun B-W, Lee I-J (2017) Chloroplast genomes of Arabidopsis halleri ssp. gemmifera and Arabidopsis lyrata ssp. petraea: Structures and comparative analysis. Scientific reports 7: 7556. https://doi.org/10.1038/s41598-017-07891-5.
4. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology 19: 455–477. https://doi.org/10.1089/cmb.2012.0021.
5. Bi G, Mao Y, Xing Q, Cao M (2018) HomBlocks: a multiple-alignment construction pipeline for organelle phylogenomics based on locally collinear block searching. Genomics 110: 18–22. https://doi.org/10.1016/j.ygeno.2017.08.001.
6. Darriba D, Taboada G, Doallo R (2012) Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nature Methods 9: 772. https://doi.org/10.1038/nmeth.2109.
7. Doorduin L, Gravendeel B, Lammers Y, Ariyurek Y, Chin-A-Woeng T, Vrieling K (2011) The complete chloroplast genome of 17 individuals of pest species Jacobaea vulgaris: SNPs, microsatellites and barcoding markers for population and phylogenetic studies. DNA research 18: 93–105. https://doi.org/10.1093/dnares/dsr002.

8. Dubchak I, Ryaboy DV (2005) VISTA family of computational tools for comparative analysis of DNA sequences and whole genomes. Methods Mol Biol 338: 69–89.
9. Gao F, Chen C, Arab DA, Du Z, He Y, Ho SY (2019) EasyCodeML: A visual tool for analysis of selection using CodeML. Ecology and Evolution 9: 3891–3898. https://doi.org/10.1002/ece3.5015.
10. Greiner S, Lehwark P, Bock R (2019) OrganellarGenomeDRAW (OGDRAW) version 1.3. 1: expanded toolkit for the graphical visualization of organellar genomes. Nucleic acids research 47: W59-W64. https://doi.org/10.1093/nar/gkz238.
11. Grey-Wilson C (2000) Clematis, the genus: A comprehensive guide for gardeners, horticulturists and botanists. (No Title):
12. Guo X, Wang G, Li J, Li J, Sun X (2023) Analysis of Floral Color Differences between Different Ecological Conditions of Clematis tangutica (Maxim.) Korsh. Molecules 28: 462. https://doi.org/10.3390/molecules28010462.
13. Huang JL, Sun GL, Zhang DM (2010) Molecular evolution and phylogeny of the angiosperm ycf2 gene. Journal of Systematics and Evolution 48: 240–248. https://doi.org/10.1111/j.1759-6831.2010.00080.x.
14. Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17: 754–755. https://doi.org/10.1093/bioinformatics/17.8.754.
15. Kanzaki N, Kiontke K, Tanaka R, Hirooka Y, Schwarz A, Müller-Reichert T, Chaudhuri J, Pires-daSilva A (2017) Description of two three-gendered nematode species in the new genus Auanema (Rhabditina) that are models for reproductive mode evolution. Scientific reports 7: 11135. https://doi.org/10.1038/s41598-017-09871-1.
16. Li Z-H, Ma X, Wang D-Y, Li Y-X, Wang C-W, Jin X-H (2019) Evolution of plastid genomes of Holcoglossum (Orchidaceae) with recent radiation. BMC Evolutionary Biology 19: 1–10. https://doi.org/10.1186/s12862-019-1384-5.
17. Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA g-enes in genomic sequence. Nucleic acids research 25: 955–964. https://doi.org/10.1093/nar/25.5.955.
18. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1: 2047-2217X-2041-2018. https://doi.org/10.1186/2047-217X-1-18.
19. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, Lanfear R (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Molecular biology and evolution 37: 1530–1534. https://doi.org/10.1093/molbev/msaa015.
20. Raveendar S, Na Y-W, Lee J-R, Shim D, Ma K-H, Lee S-Y, Chung J-W (2015) The complete chloroplast genome of Capsicum annuum var. glabriusculum using Illumina sequencing. Molecules 20: 13080–13088. https://doi.org/10.3390/molecules200713080.
21. Shaw J, Lickey EB, Schilling EE, Small RL (2007) Comparison of whole chloroplast genome se-quences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. American journal of botany 94: 275–288. https://doi.org/10.3732/ajb.94.3.275.
22. Shi C, Hu N, Huang H, Gao J, Zhao Y-J, Gao L-Z (2012) An improved chloroplast DNA extract-ion procedure for whole plastid genome sequencing. PLoS One 7: e31468. https://doi.org/10.1371/journal.pone.0031468.
23. Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-Shinozaki K (1986) The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. The EMBO journal 5: 2043–2049. https://doi.org/10.1002/j.1460-2075.1986.tb04464.x.

24. Wang Y, Zhan D-F, Jia X, Mei W-L, Dai H-F, Chen X-T, Peng S-Q (2016) Complete chloroplast genome sequence of Aquilaria sinensis (Lour.) Gilg and evolution analysis within the Malvales order. Front Plant Sci 7: 280. https://doi.org/10.3389/fpls.2016.00280.
25. Wen-Tsai W, Liang-Qian L (2005) A new system of classification of the genus Clematis (Ranun-culaceae). Journal of Systematics and Evolution 43: 431. https://doi.org/10.1360/aps040130.
26. Wicke S, Schneeweiss GM, Depamphilis CW, Müller KF, Quandt D (2011) The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant molecular biology 76: 273–297. https://doi.org/10.1007/s11103-011-9762-4.
27. Xu B, Yang Z (2013) PAMLX: a graphical user interface for PAML. Molecular biology and evolution 30: 2723–2724. https://doi.org/10.1093/molbev/mst179.
28. Yang M, Zhang X, Liu G, Yin Y, Chen K, Yun Q, Zhao D, Al-Mssallem IS, Yu J (2010) The complete chloroplast genome sequence of date palm (Phoenix dactylifera L.). PLoS One 5: e12762. https://doi.org/10.1371/journal.pone.0012762.
29. Yang Y, Yuanye D, Qing L, Jinjian L, Xiwen L, Yitao W (2014) Complete Chloroplast Genome Sequence of Poisonous and Medicinal Plant Datura stramonium: Organizations and Implicatio-ns for Genetic Engineering. PLoS One 9: e110656. https://doi.org/10.1371/journal.pone.0110656.
30. Zhang R, Zhang Z, Hou Z, Li C, Han D, Xie Z (2006) The discussion of technology of flavonoiods extraction from flowers of Clematis tangutica (Maxim.) Korsh. Chinese Wild Plant Resources 25: 58–60.
31. Zhao M, Da-Wa Z-M, Guo D-L, Fang D-M, Chen X-Z, Xu H-X, Gu Y-C, Xia B, Chen L, Ding L-S (2016) Cytotoxic triterpenoid saponins from Clematis tangutica. Phytochemistry 130: 228–237. https://doi.org/10.1016/j.phytochem.2016.05.009.

Figures

Figure 1

Clematis. tangutica (Maxim.) Korsh.

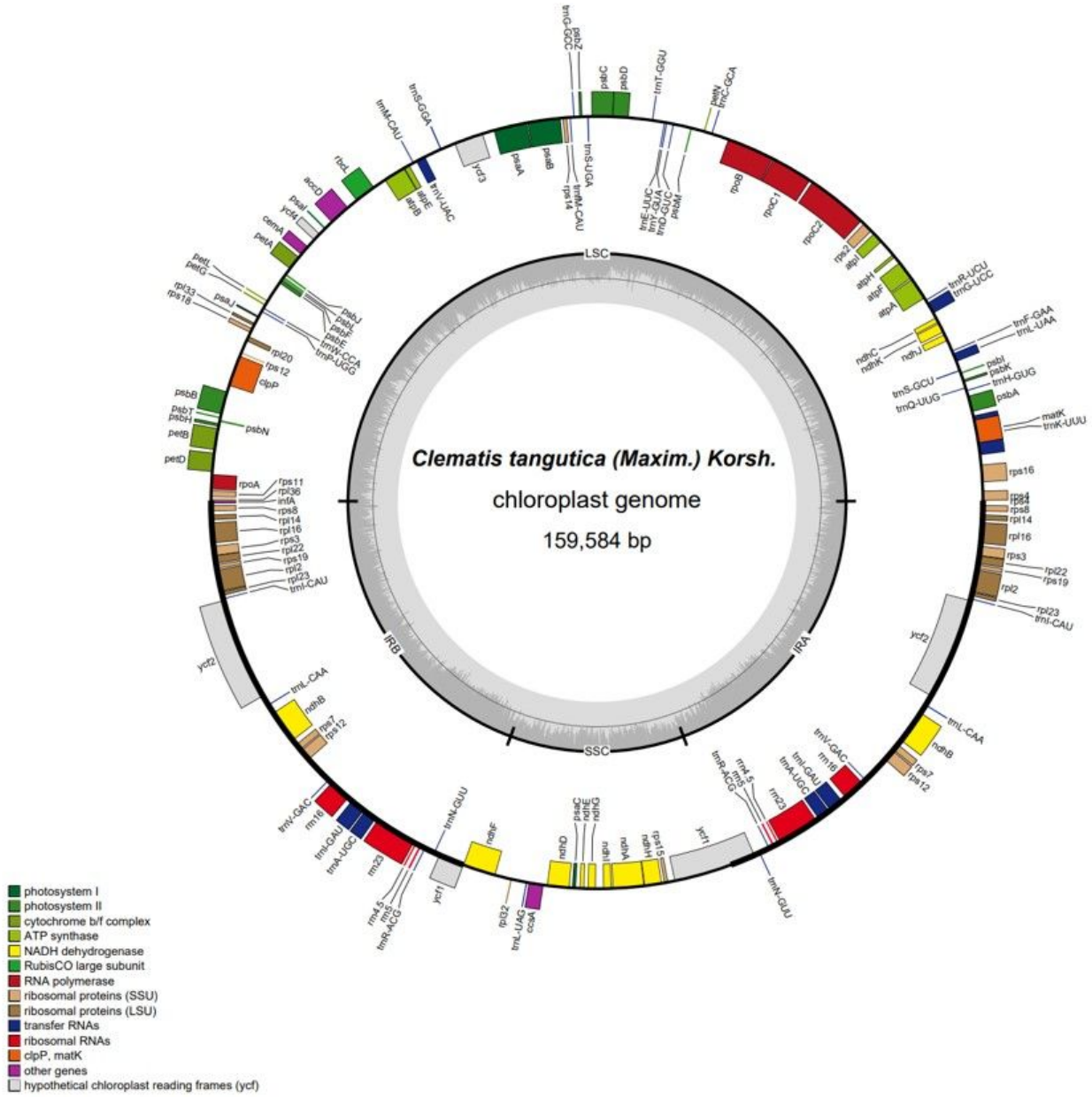

Figure 2

Gene map of the *C. tangutica* (Maxim.) Korsh. chloroplast genome. Genes drawn outside of the circle are transcribed counter-clockwise, while genes shown on the inside of the circle are transcribed clockwise. Genes belonging to different functional groups are color-coded. The darker gray in the inner circle indicates GC content, while the lighter gray corresponds to AT content and the gray line in the middle is the 50% threshold line.

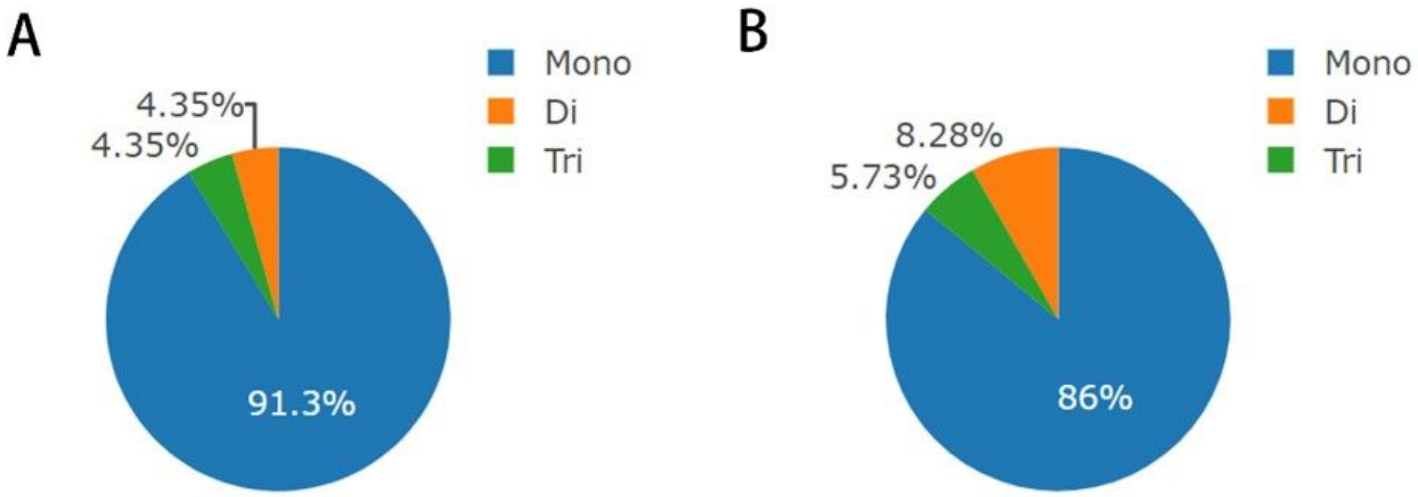

Figure 3

A.SSR counts for each type of distribution; B. SSR length distribution per type

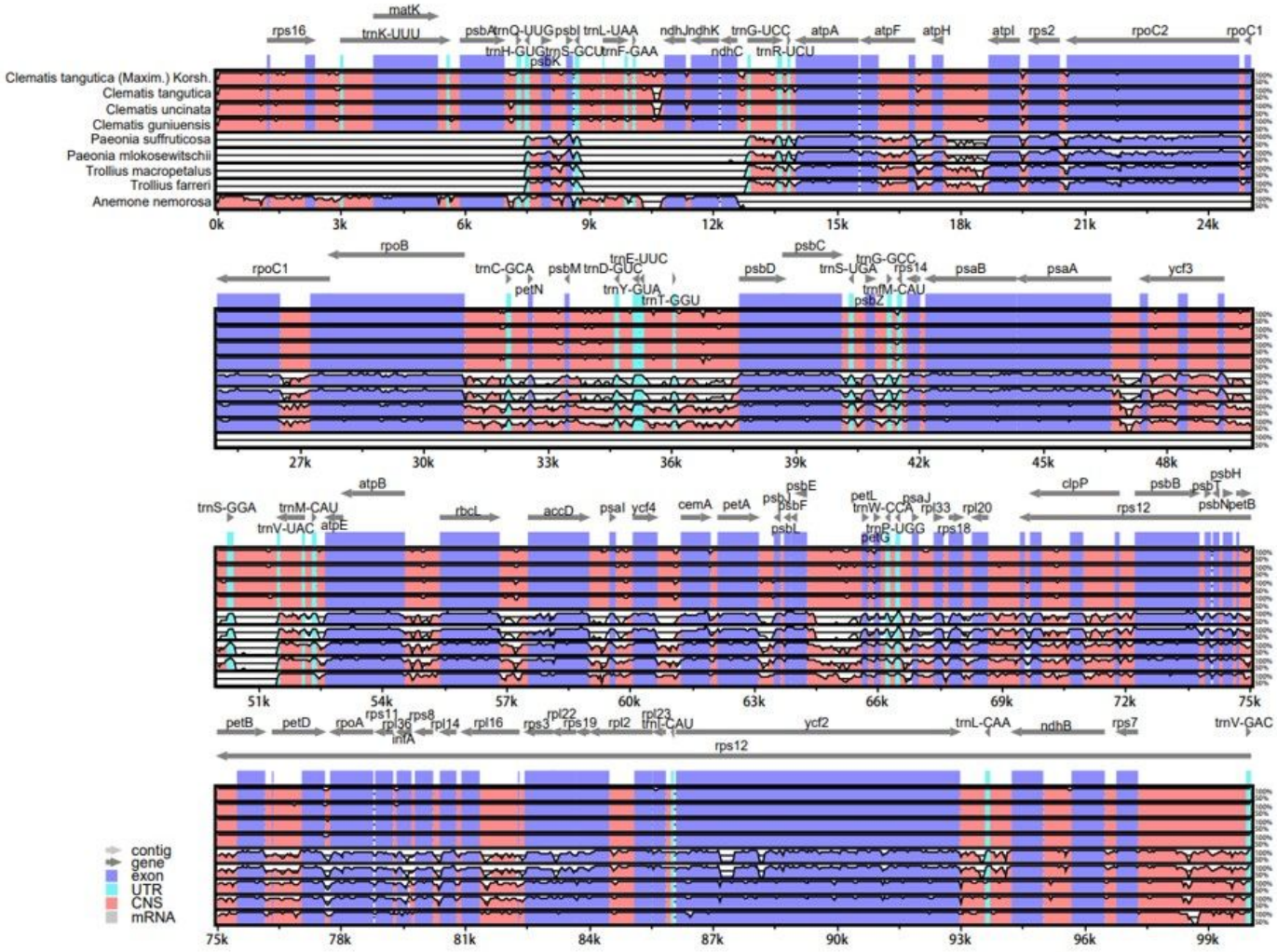

Figure 4

Percent identity plot for the comparison of 9 Ranunculaceae chloroplast genomes. The whole chloroplast genome was divided into four parts, and the gene names are displayed in sequence on the top line of each part (arrows indicate the transcriptional direction). The sequence similarity of the alignment region of *C. tangutica* (Maxim.) Korsh. and seven other species is shown as the filling color in each black stripe. The x-axis indicates the position of the chloroplast genome at a certain site, and the y-axis indicates the average sequence identity percentage (50–100%) with *C. tangutica* (Maxim.) Korsh. on the position of a species at a certain position (50–100%). The coding sequences (exons), rRNA, tRNA and the conserved non-coding sequences (CNS) in the genomic region are represented with different colors.

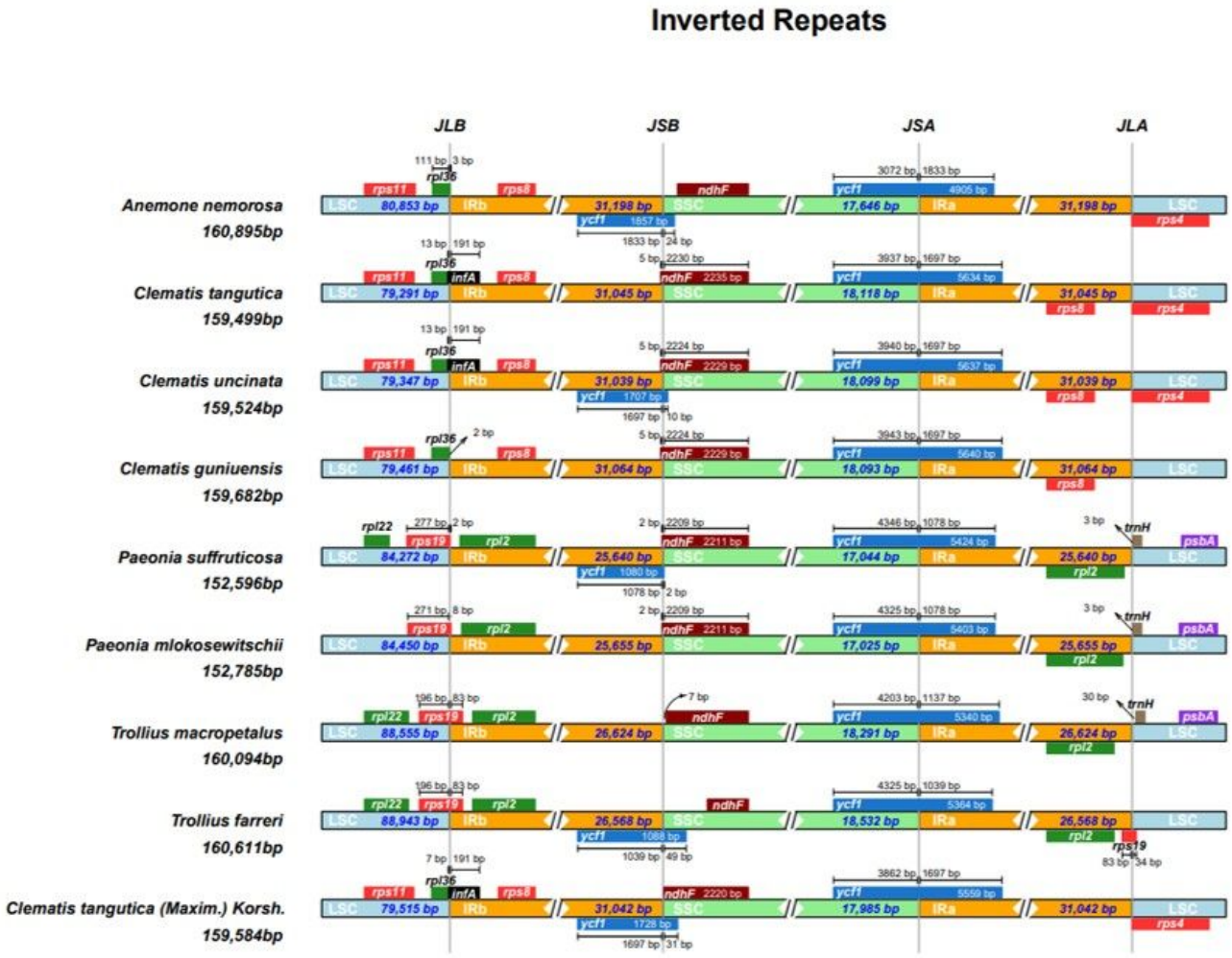

Figure 5

Comparison of the borders of the LSC, SSC and IR regions among the nine chloroplast genomes.

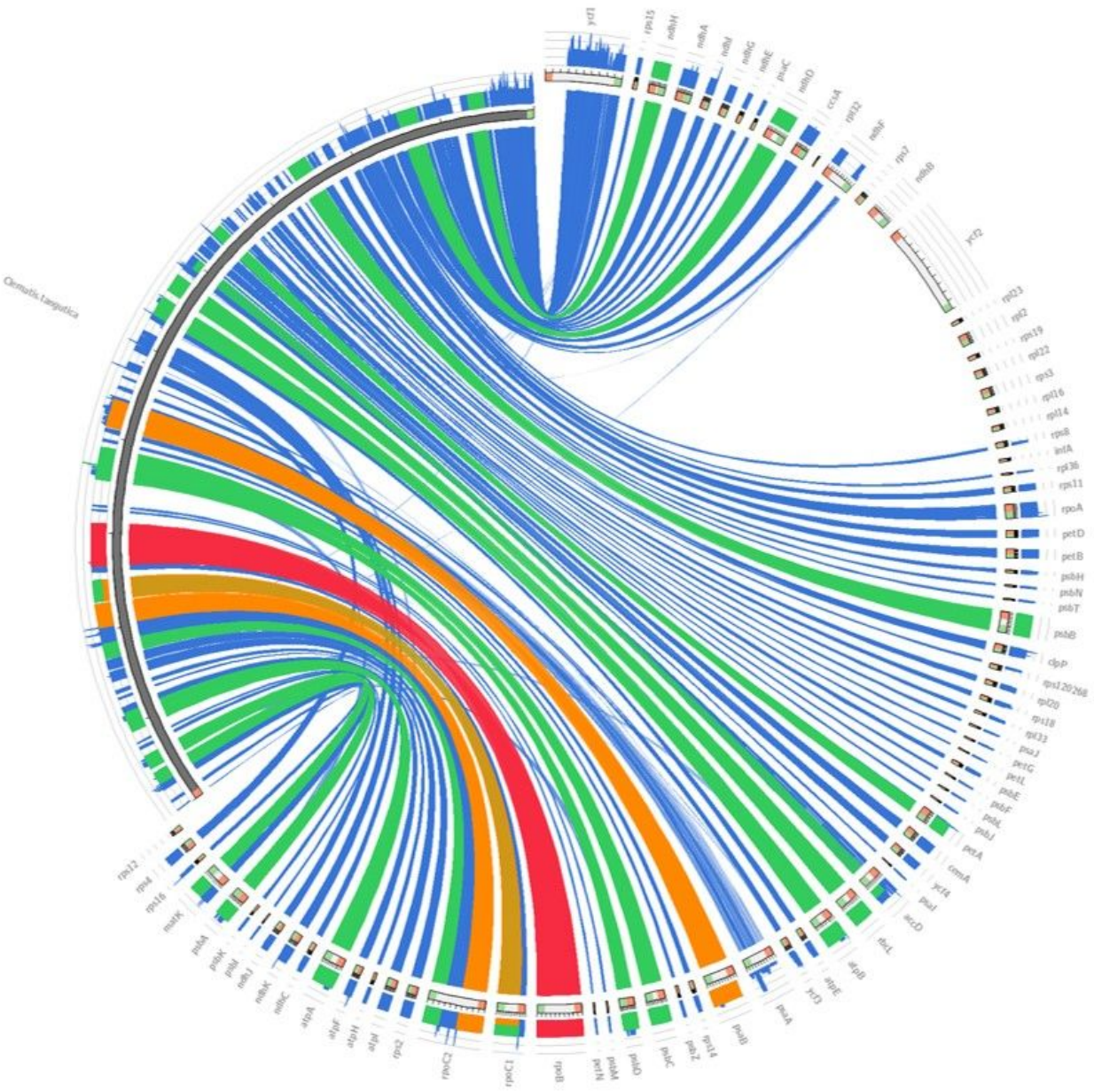

Figure 6

Comparison of the similarity of chloroplast genomes between *C. tangutica* (Maxim.) Korsh. and eight other species of crops in the Ranunculaceae.

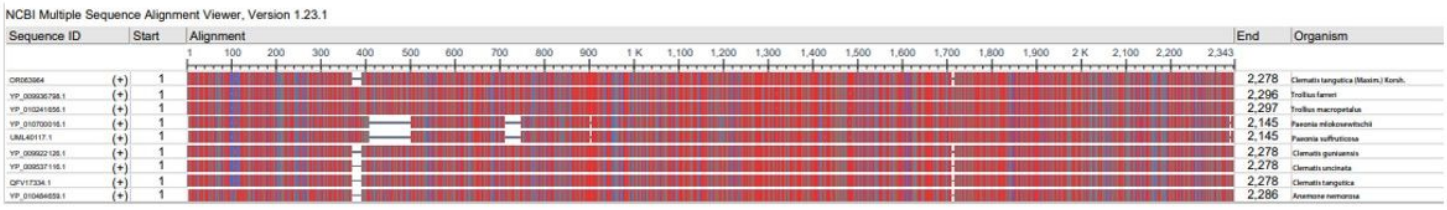

Figure 7

Comparison of the *ycf2* gene sequence in chloroplast genomes between *C. tangutica* (Maxim.) Korsh. and eight other species of crops in the Ranunculaceae. The white vacancy corresponds to the missing amino acid sequence.

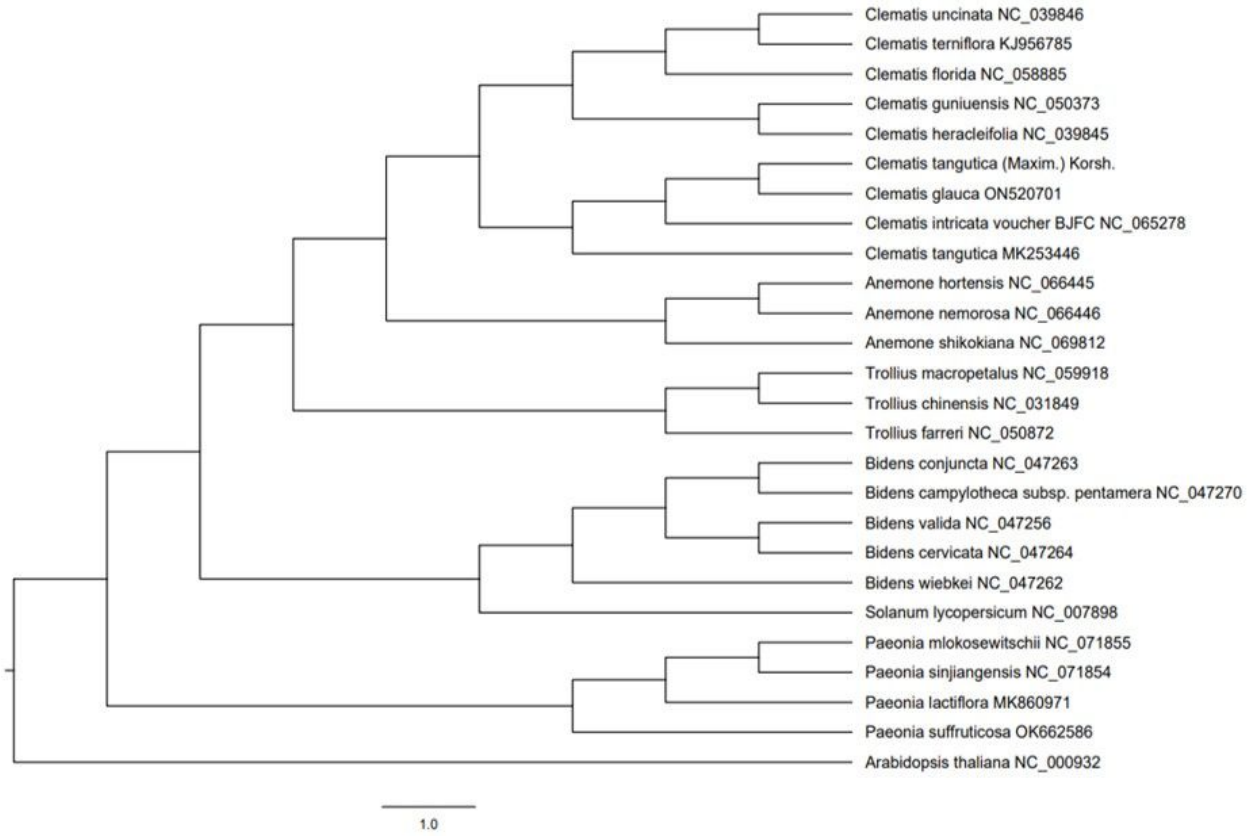

Figure 8

Molecular phylogenetic tree of 26 species based on a neighbor joining analysis.

Numbers above and below nodes are bootstrap support values 50%.