Journal of Applied Botany and Food Quality 95, 51 - 59 (2022), DOI:10.5073/JABFQ.2022.095.007

Key lab of Chinese Medicine Resources Conservation, State Administration of Traditional Chinese Medicine of China, Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing 100193, China

The complete chloroplast genome sequence of *Gentiana triflora* and comparative analysis with its congeneric species

Xinke Zhang, Yujing Miao, Xiao Sun, Yuan Jiang, Tiexin Zeng, Yan Zheng, LinFang Huang*

(Submitted: October 30, 2021; Accepted: March 4, 2022)

Summary

Gentiana triflora is an important medicinal plant in China with economic and medicinal value. Here, we report the complete chloroplast sequences of *G. trifloral*. The cp genome of *G. triflora* of 149, 125 bp contains 130 unique genes, including 85 protein-coding genes, 8 rRNA genes, and 37 tRNA genes. The analysis of repeat showed that palindromic had the highest frequency. Besides, a total number of 45 SSR were identified, most of which were mononucleotide adenine-thymine. Comparative genome analysis of *Gentiana* species revealed that the pair of the inverted repeat was more conserved than the single-copy region. This analysis resulted in identification of 8 hypervariable regions *(trnH*-GUG, *trnG*-UCC-intron, *atp*I, *trnD*-GUC, *trnL*-UAA, *rpl*32-*trnL*-UAG, *pet*A and *ycf*1). Phylogenetic analysis revealed that *G. triflora* was most closely related to *Gentiana manshurica*. In conclusion, this study enriched the genomic resources of the *Gentiana* genus and provided a basis for evolution and phylogeny analyses.

Keywords: *Gentiana triflora,* Chloroplast genome, Comparative analysis, Phylogenetic relationship

Introduction

The genus *Gentiana* is a major group in the Gentianaceae family, comprising around 400 species that are widespread across the Northwest of Africa, America, Europe, Asia, and East of Australia, of which about 247 species are found in China (YUAN et al., 1996). The genus *Gentiana* has high medicinal values. Especially the roots of sect. *Pneumonanthe* and sect. *Cruciata* are widely used as a remedy in traditional Chinese medicine for more than 2000 years.

Gentiana triflora is the perennial herb belonging to sect. *Pneumonanthe* with brilliant blue flowers, which has high ornamental value. Besides, it can be used as medicine and officially listed as "Longdan" in the Chinese Pharmacopoeia, along with *G. manshurica, G. scabra,* and *G. rigescens*. The main effective constituent is gentiopicroside, which has hepatoprotective, analgesic, antioxidation, and antitumor effects (SONG et al., 1987; WANG et al., 2004). To date, there have been many studies on the pharmacological activity and chemical composition of *G. triflora* (YAMADA et al., 2014; DOI et al., 2010). However, few data are available regarding genomic studies. Previous studies have used DNA fragments to explore the phylogenetic position of *G. triflora* in the *Gentiana* genus. Compared with the complete chloroplast (cp) genome, DNA fragments provided limited genetic information, which fail to distinguish closely related species and construct accurate phylogenetic relationships (MISHIBA et al., 2009).

In plant cell, the chloroplast (cp) is a photosynthetic organelle that is associated with metabolism like starch, amino acids, pigments, and fatty acids (NEUHAUS and EMES, 2000). Plant cp genome is highly conserved, being circular with a length of 120-170 kb, consisting of four typical areas: large single-copy(LSC) and small single-copy(SSC) regions, and two inverted repeats(IRs). The size of LSC ranges from 80 to 90 kb, while SSC ranges from 16 to 27 kb. The LSC and SSC are harbor genes related to Photosystem I(*psa*) and Photosystem II(*psb*), which are separated by the IRs. The two IRs are between 20 and 30 kb in size and have the same sequences but in the opposite direction. The length of cp genome differences are mostly due to the expansion, contraction, and loss of the IRs (PALMER, 1997; WICKE et al., 2011). It is generally believed to contain 120-130 genes, including of 80-90 protein-coding genes, 8 rRNA genes, and approximately 30 tRNA genes (ALDRIDGE et al., 2005).

With the development of next-generation sequencing, many cp genomes in the genus *Gentiana*, have been sequenced. To date, three species of the original plant of Longdan cp genomes have been related studies except for *G. triflora*, which range from 146915 to 149185 bp (SHANG et al., 2020; YANG et al., 2020; LIANG et al., 2020). In this research, we sequenced and analyzed the cp genome of *G. triflora*, and analyzed its general characteristics, IR contraction and expansion, codon usage, and simple sequence repeats (SSRs). We compared *G. triflora* with other plants in the *Gentiana* genus in terms of gene content, organization, and phylogenetic relationships. The result of this study will provide resources for evolutionary studies and authentication of *G. triflora*, and the reference for research on phylogenetic relationships within the *Gentiana* genus.

Materials and methods

Plant material and DNA extraction

The fresh leaves of *G. triflora* were collected from Qixing Lake Wetland Park (Hebei China, 41°35' N, 116°32' E). The samples were identified by Professor Linfang Huang and stored at the Herbarium of the Chinese Academy of Medical Science and Peking Union Medicinal College (CMPB12686, CMPB12687and CMPB12688). The total DNA was isolated by modified cetyl trimethyl ammonium bromide (CTAB) method (ALLEN et al., 2006), and paired-end reads were generated by using Illumina NovaSeq system (Illumina, San Diego, CA, USA).

Genome assembly and annotation

Raw data reads were filtered by GetOrganelle pipeline (https://github.com/Kinggerm/GetOrganelle) to produce clean data (JIN et al., 2020). Then, filtered reads were assembled with NOVOPlasty version 3.8.3 (DIERCKXSENS et al., 2017). The genome was automatically annotated using CpGAVAS2 (SHI et al., 2019). Further, using Apollo (LEWIS et al., 2002) to correct the annotations with problems, and genome map was drawn by OGDRAW (GREINER et al., 2019). The sequence and annotation of the genome have been submitted to GenBank with accession number MZ087943.

Repeats and codon usage analysis

Repetitive sequences analysis and codon usage were performed using CPGAVAS2 analysis. The SSR including mono-, di-, tri-,

* Corresponding author

tetra-, pentamore-, and hexa-nucleotides, were identified using MISA (https://webblast.ipkgatersleben.de/misa/) (BEIER et al., 2017), and the minimum numbers were set as 10, 5, 4, 3, 3, and 3, respectively. The size and location of repeats sequences, including forward, reverse, palindromic and complement repeats were identified by running the REPuter (KURTZ et al., 2001).

Phylosuite (ZHANG et al., 2020) was used to extract the protein-coding genes, and then CodonW version 1.4.2 (SHARP et al., 1986) was used to analyze codon usage and relative synonymous codon usage (RSCU) values.

Prediction of RNA-editing sites

The predictive RNA Editor for Plants (PREP) (MOWER, 2005) was used to predict potential RNA editing sites in the protein-coding genes of cp genome with the cutoff value of 0.8.

Genome comparison

The mVISTA program was used to compare the cp genomes of the 23 *Gentiana* species and determine interspecific variations (http://genome.lbl.gov/vista/mvista/submit.shtml (FRAZER et al., 2004). The 64 homologous protein-coding genes from the 23 species were identified and extracted using PhyloSuite (ZHANG et al., 2020). The nucleotide sequences were aligned using MAFFT (v7.450) (JOHN et al., 2019). MEGA v6.0 (TAMURA et al., 2013) was used to calculate the percentage of variable positions in the protein-coding genes. Then, DnaSP v6.0 with 500 bp window length and 500 bp step length was used to calculate the nucleotide diversity (Pi) (ROZAZ, 2009). Finally, we used IRscope (https://irscope.shinyapps.io/irapp/) to compare the IR expansion and contraction of cp genomes among *Gentiana* genus (AMIRYOUSEFI et al., 2018).

Analysis of the nucleotide substitution rate

The CODEML module in PAML v4.9 (MOWER, 2009) was used to estimate rates of nucleotide substitution, including nonsynonymous (dN), synonymous (dS), and the ratio of nonsynonymous to synonymous rates (dN/dS) (YANG, 2007). The detailed parameters were as follows: CodonFreq = 2 (F3 × 4 model); model = 0 (allowing a single dN/dS value to vary among branches); clean data = 1 (removing sites with ambiguous data); and other parameters in the CODEML were left to their default values. Then, a phylogenetic tree for each gene was generated using the maximum likelihood (ML) method implemented in RAxML v8.2.4 (STAMATAKIS, 2014).

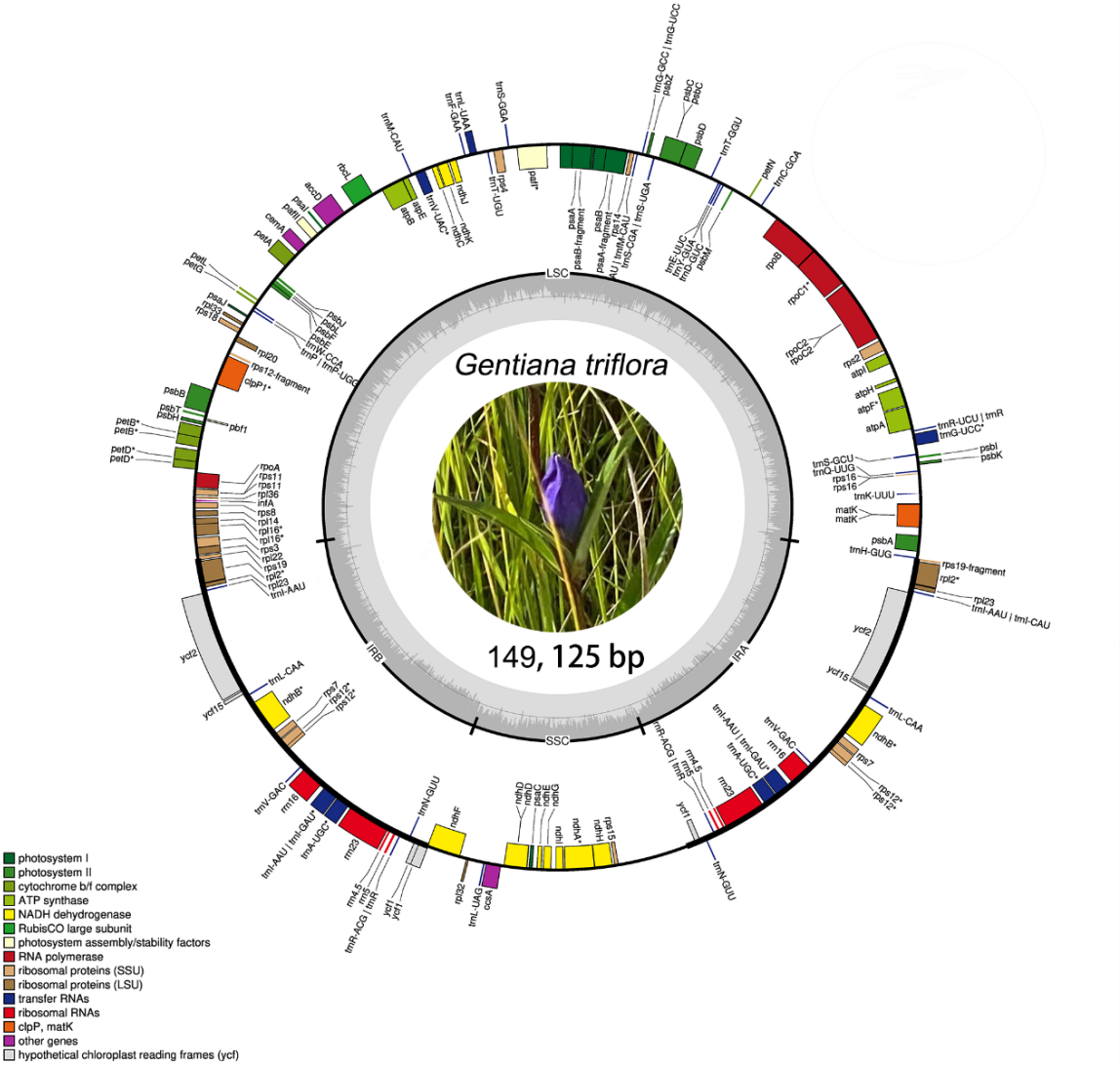


Fig. 1: Gene map of *G. triflora* chloroplast genome. Thick lines represent inverted repeats (IRa and IRb) which divide the genome into large and small single-copy regions (LSC and SSC). Features on the clockwise- and counter-clockwise-transcribed strands are drawn on respectively the inside and outside of the circle. The genes belonging to different functional categories were color-coded, which are shown in the left corner.

Phylogenetic analysis

The 22 complete cp genomes among the *Gentiana* genus were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/). *Comastoma pulmonarium* (NC_050654.1) was used as an outgroup. The cp genome sequences were aligned using MAFFT (v7.450) (John et al., 2019). The phylogenetic tree was constructed using the maximum likelihood (ML) method implemented in RAxML v8.2.4 (Stamatakis, 2014). The parameters were set to "raxmlHPC-PTHREADS-SSE3 -f a -N 1000 -m GTRGAMMA -x 551314260 -p 551314260". The bootstrap analysis was performed with 1000 replicates.

Results

Characteristics of *G. triflora* chloroplast genome

The complete cp genome size of *G. triflora* was reported to be 149, 125 bp in size and has a structural organization of quadripartite including a pair of IRs, a LSC and SSC with 25, 262 bp, 81, 316 bp and 17, 285 bp, respectively (Fig. 1 and Tab. 1). The GC content analysis showed that the overall GC content was 37.61%. The IR regions have higher GC content of 43.37% compared with the SSC and LSC regions with 31.27% and 35.38% respectively (Tab. 1). The overall GC content was lower than AT content, which is a feature generally observed in cp genomes of angiosperm plants.

Genome annotation

In total, 130 unique genes were identified in the *G. triflora* cp genome, including 85 protein-coding genes, 37 tRNA genes, and 8 rRNA genes (Tab. 2). Most genes could be divided into three groups, respectively self-replication, photosynthesis, and other genes (Tab. 2). Introns play an important role in protein-coding gene splicing (Fu et al., 2021). Seven protein-coding genes (*ndh*A, *ndh*B, *rpl*16, *rpl*2, *atp*F, *pet*B, *rpo*C1) and five tRNA (*trnA*-UGC, *trnV*-UAC, *trnG*-UCC, *trnK*-UUU, *trnL*-UAA) contained only one intron, while three genes (*pet*D, *clp*P, *ycf*3) contained two introns. Seven protein-coding genes, six tRNA genes, and four rRNA genes were completely duplicated in the IR regions, which exist as two copies.

Repetitive sequence and codon usage analysis

In this study, a total of 45 SSRs were detected in the cp genome of *G. triflora* (Fig. 2a). Among all SSRs, mononucleotide repeats were the most common, of which A/T accounts for 75.56% of the total. However, no pentanucleotide and hexanucleotide were observed in

Tab. 1: Basic features of chloroplast genome of *G. triflora*.

Species		*G. triflora*
Accession number		MZ087943
Length (bp)	Total	149 .125
	LSC	81 .316
	SSC	17 .285
	IR	25 .262
GC content (%)	Total	37 .61
	LSC	35 .38
	SSC	31 .27
	IR	43 .37
Gene numbers	Total	116
	Protein-coding gene	78
	tRNA gene	30
	rRNA gene	8

Tab. 2: Gene composition of *G. triflora* chloroplast genome.

Category of genes	Group of genes	Name of genes
RNA	rRNA	*rrn*16S[c], *rrn*23S[c], *rrn*4.5S[c], *rrn*5S[c]
		trnA-UGC[a,c], *trnC*-GCA, *trnD*-GUC, *trnE*-UUC, *trnF*-GAA, *trnfM*-CAU, *trnG*-GCC, *trnG*-UCC[a], *trnH*-GUG, *trnI*-CAU[c], *trnI*-GAU[a,c], *trnK*-UUU[a], *trnL*-CAA[c], *trnL*-UAA[a], *trnL*-UAG, *trnM*-CAU, *trnN*-GUU[c], *trnP*-UGG, *trnQ*-UUG, *trnR*-ACG[c], *trnR*-UCU, *trnS*-GCU, *trnS*-GGA, *trnS*-UGA, *trnT*-GGU, *trnT*-UGU, *trnV*-GAC,
	tRNA	*trnV*-UAC[a], *trnW*-CCA, *trnY*-GUA
Genes for photosynthesis	Subunits of ATP synthase	*atp*A, *atp*B, *atp*E, *atp*F[a], *atp*H, *atp*I
	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[b]
	Subunits of NADH dehydrogenase	*ndh*A[a], *ndh*B[a,c], *ndh*C, *ndh*D, *ndh*E, *ndh*F, *ndh*G, *ndh*H, *ndh*I, *ndh*J, *ndh*K
	Subunits of cytochrome b/f complex	*pet*A, *pet*B[a], *pet*D[b], *pet*G, *pet*L, *pet*N
	Subunits of photosystem I	*psa*A, *psa*B, *psa*C, *psa*I, *psa*J
	Subunit of rubisco	*rbc*L
Self-replication	Large subunit of ribosome	*rpl*14, *rpl*16[a], *rpl*2[a,c], *rpl*20, *rpl*22, *rpl*23[c], *rpl*32, *rpl*33, *rpl*36
	DNA dependent RNA polymerase	*rpo*A, *rpo*B, *rpo*C1[a], *rpo*C2
	Small subunit of ribosome	*rps*11, *rps*12[c], *rps*14, *rps*15, *rps*18, *rps*19, *rps*2, *rps*3, *rps*4, *rps*7[c], *rps*8
Other genes	Subunit of Acetyl-CoA-carboxylase	*acc*D
	c-type cytochrom synthesis gene	*ccs*A
	Envelop membrane protein	*cem*A
	Protease	*clp*P[b]
	Maturase	*mat*K
Unkown	Conserved open reading frames	*ycf*1, *ycf*15[c], *ycf*2[c], *ycf*4

[a] Gene with one intron, [b] Gene with two intron and [c] Gene with two copies

this cp genome. In addition, four types of interspersed repeats were detected, comprising of 22 palindromic repeats, 20 forward repeats, 5 reverse repeats, and 2 complement repeats (Fig. 2b).

The total sequence length of the protein-coding genes for codon analysis were 78, 513 bp in *G. triflora*. These codons encoded 21 amino acids (excluding termination codons), of which Leu coded the most (5039) and Trp coded the least (671). RSCU, as a measure of nonuniform synonymous codon usage in coding sequences, is the ratio of the frequency of usage of a given codon to the expected frequency. If RSCU value is greater than 1, specific codon frequency is higher than other synonymous codons, and codon usage is used more often than expected (SHARP and LI, 1987). The RSCU values of each codon were shown in Fig. 3. Specifically, there were 36 codons used frequently with RSCU > 1. Due to the structure of the A/T-rich cp genome structure, the preferred codons ended with purines (A/U), except for UUG and GCC.

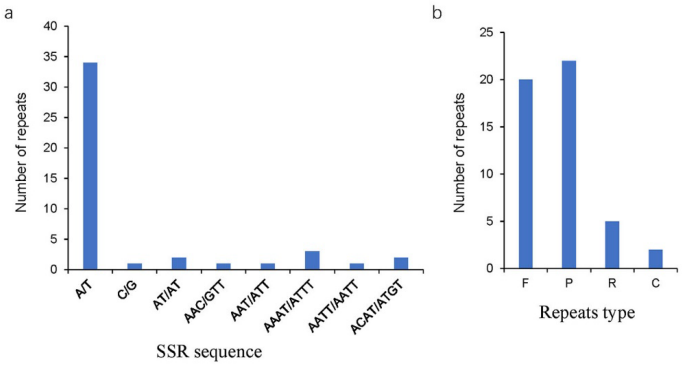


Fig. 2: Comparison of the repeats in the chloroplast genome of *G. triflora*. a. Types and numbers of SSRs detected in *G. triflora*; b. Types and numbers of interspersed repeats in *G. triflora*.
F = forward, P = palindromic, R = reverse and C = complement

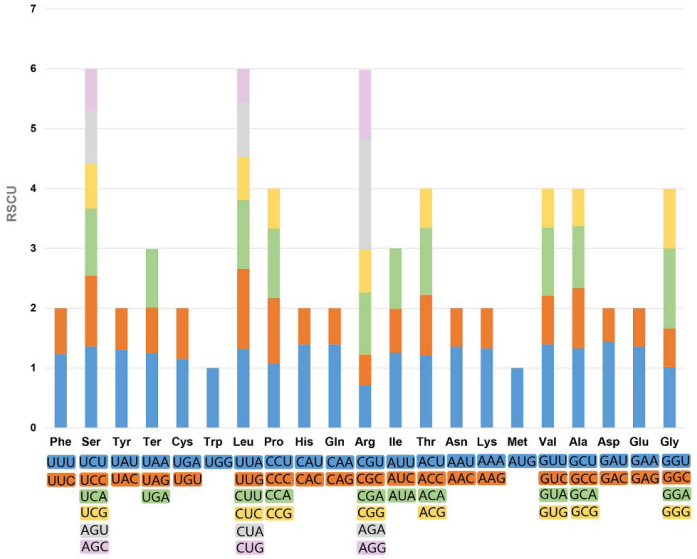


Fig. 3: The RSCU values of all protein-coding genes in the chloroplast genome of *G. triflora*.

There were overall 56 RNA editing sites in 18 genes of the cp genome of *G. triflora*. The *rpoC*2 was found to have the highest number of editing site (10 sites), and *atp*I, *mat*K, *ndh*G, *pet*B, *psa*I, *rpl*20, *rps*2 have the lowest number of editing site with 1 editing site each. All editing sites were cytosine to uracil (C-U) conversion, most of which resided in the first and second bases of codons, while, no editing sites were found at the third codon position. Among these sites, the most frequent amino acid conversion was serine to leucine (S-L). In addition, many RNA editing sites have the potential to amino acid changes, such as phenylalanine (F), leucine (L), tyrosine (Y), methionine(M), tryptophan(W), isoleucine(I) and valine (V) (Supplementary Tab. 1).

Genomic divergence

The comprehensive sequence divergence of *Gentiana* genus was analyzed using mVISTA with the annotation of *G. straminea* as a reference (DUBCHAK and RYABOY, 2005). The patterns of similarity of each cp genome region are shown in Supplement Fig. 1. Variability in the IR regions exhibited considerably lower than the LSC and SSC regions. Furthermore, although most of the protein-coding genes were conserved, there were still some protein-coding genes with large variations, such as *ndh*F and *pet*D. The variation was mainly concentrated in intergenic regions, such as *atp*H-*atp*I, *pet*N-*psb*M, *trnH*-GUG-*psb*A, *trnT*-UGU-*trnL*-GAA, *ycf*3-*trnS*-GGA and *rpl*32-*trnL*-UAG.

The expansion and contraction of IR regions borders are common evolutionary events and contribute to variation of cp genome size (BOUDREAU and TURMEL, 1995). In this study, we compared the differences between the IR and SC boundaries of *Gentiana* genus and found several genes spanning or near the boundary of the IR and SC regions (Fig. 4). The *rps19* gene was located in the boundary regions between LSC/IRb of *Gentiana* genus except for *G. macrophylla* and the *trn*H gene was at the border of SSC/IRb except for *G. macrophylla* and *G. apiata*. In 16 species, the *ndh*F gene had 49-54 bp in the IRb region and 2,175-2,186 bp in the SSC region. At the SSC/IRb border, *ycf*1 was a critical gene that spans the IRb region and the SSC region. Partial *ycf*1 was located in the IRb/SSC border, and complete *ycf*1 was located in the IR region at the SSC/IRa border.

To understand the genetic diversity of *Gentiana* genus, analyses were conducted using Dna-SP v6.0, and we detected 8 hypervariable regions with Pi values over 0.04 (Fig. 5), including *trnH*-GUG-*psb*A (Pi = 0.04515); *trnG*-UCC-intron (Pi = 0.04236); *atp*I (Pi = 0.04079); *trnD*-GUC (Pi = 0.04610); *trnL*-UAA (Pi = 0.04229); *rpl*32-*trnL*-UAG (Pi = 0.06800); *pet*-A (Pi = 0.04468), and most regions of the *ycf*1 gene (Pi values ranging from 0.04140 to 0.05081). Notably, most regions of the cp genome sequences had Pi values greater than 0.02 (except for IR regions), indicating abundant polymorphism of the cp genome in *Gentiana* genus.

Divergence of protein-coding genes

The rates of nonsynonymous (dN), synonymous (dS) and dN/dS ratio were calculated to detect the selective pressure among protein-coding genes of *Gentiana* genus (Fig. 6). The dN/dS of most genes was less than 0.6), suggesting that they might be under purifying selection, while, seven genes (*cem*A, *mat*K, *pet*L, *psb*H, *rpl*16, *rpl*20, and y*cf*2) had higher dN/dS ratios which ranged from 0.6 to 1.0. In addition, the dN/dS ratios were mostly less than 1.0, indicating strong purifying selection.

Phylogenetic analysis

To construct maximum likelihood (ML) tree, we used the complete cp genome sequences. The phylogenetic tree was rooted with *Comastoma pulmonarium* as outgroup and all nodes had highest bootstrap support (Fig. 7). According to phylogenetic analysis, the *Gentiana* genus could be grouped into two clades. Twelve species (*G. siphonantha*, *G. officinalis*, *G. macrophylla*, etc.) belong to one clade, and the others belong to another clade. The *G. triflora* was closely related to the *G. manshurica*.

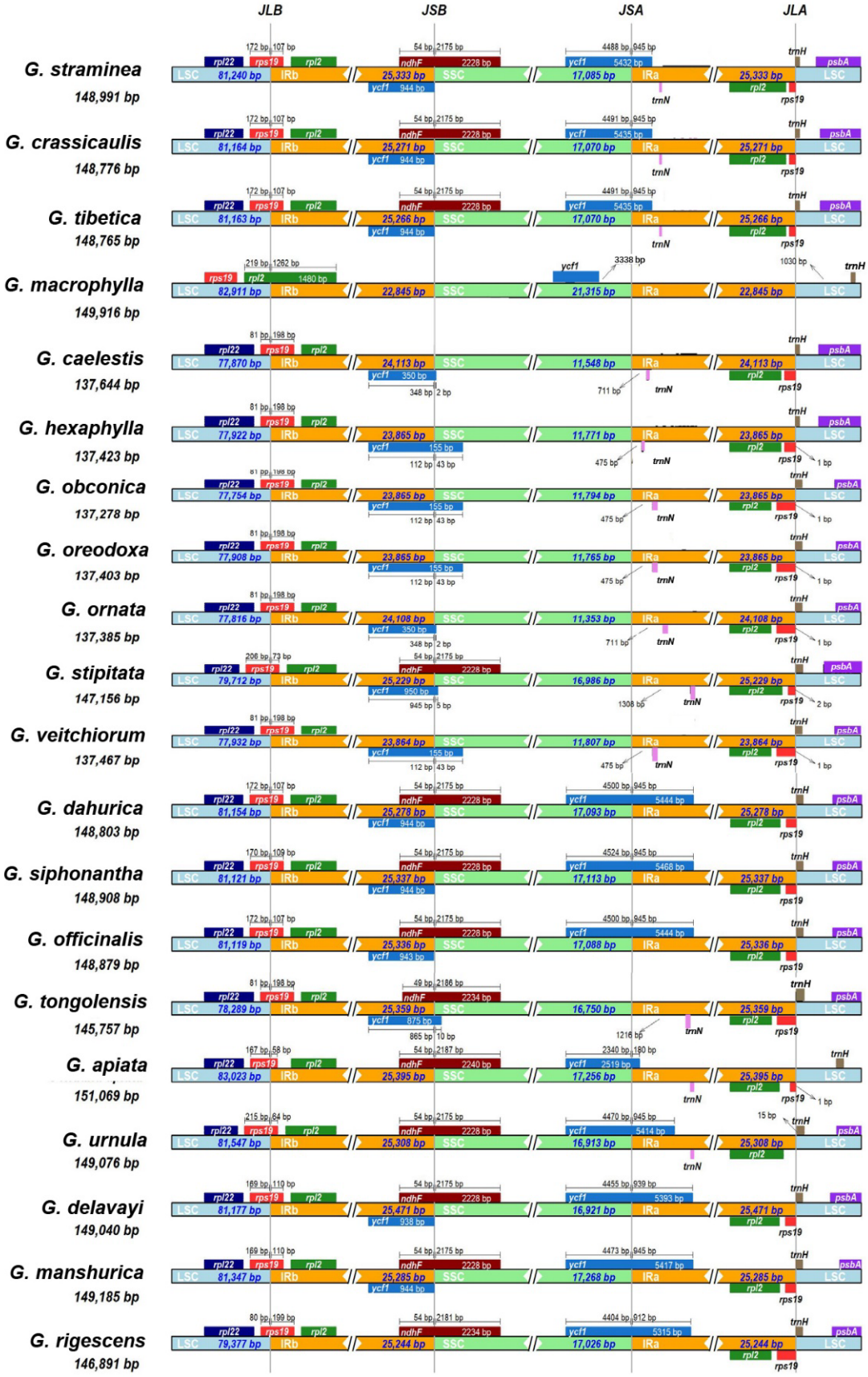


Fig. 4: Comparison of the borders among the LSC, SSC, and IR regions of *Gentiana*. JLB, JSB, JSA, and JLA indicate junction sites of LSC/IRb, IRb/SSC, SSC/IRa, and IRa/LSC, respectively. The genes around the borders are shown above or below the main line. Light blue color refers to LSC region, light green color refers to SSC region, and saffron yellow refers to a pair of IR regions.

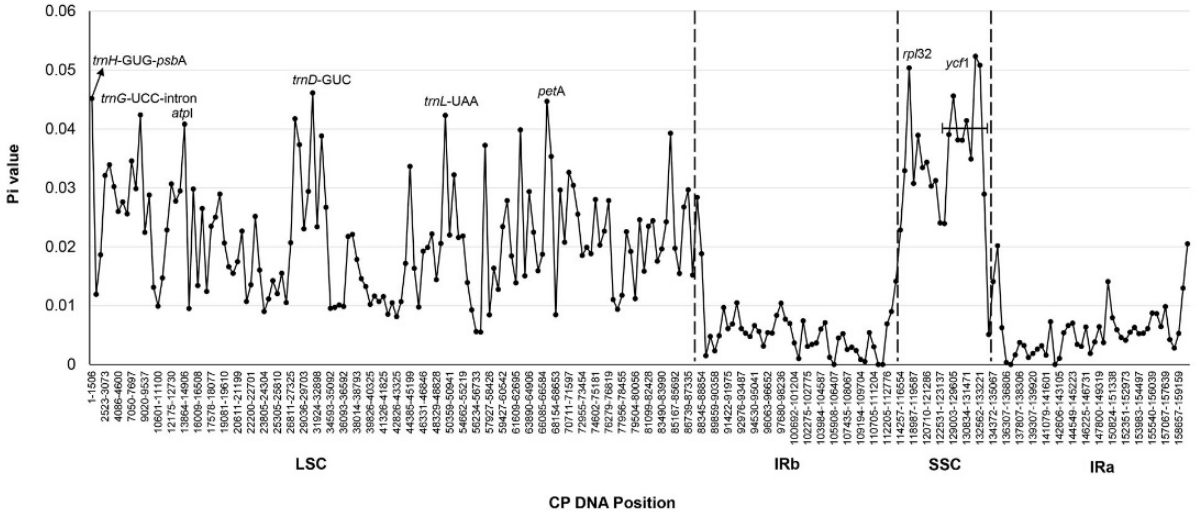


Fig. 5: Nucleotide diversity (Pi) of chloroplast genome among the *Gentiana* species. Each black dot represents the nucleotide diversity per 500 bp. LSC: large single-copy region; SSC: small single-copy region; IRa and IRb: inverted repeats.

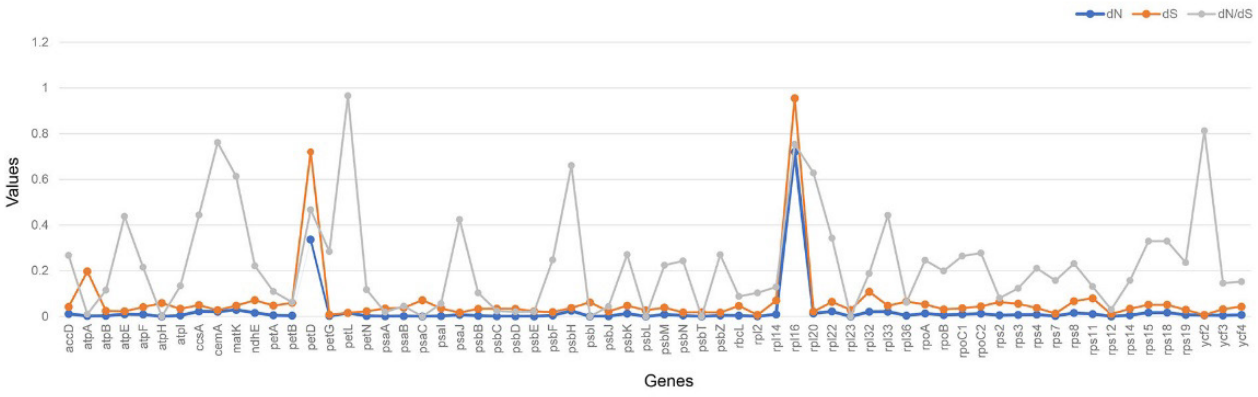


Fig. 6: The non-synonymous (dN), synonymous (dS) and dN/dS ratio values of 64 protein-coding genes of chloroplast genome among the *Gentiana* species.

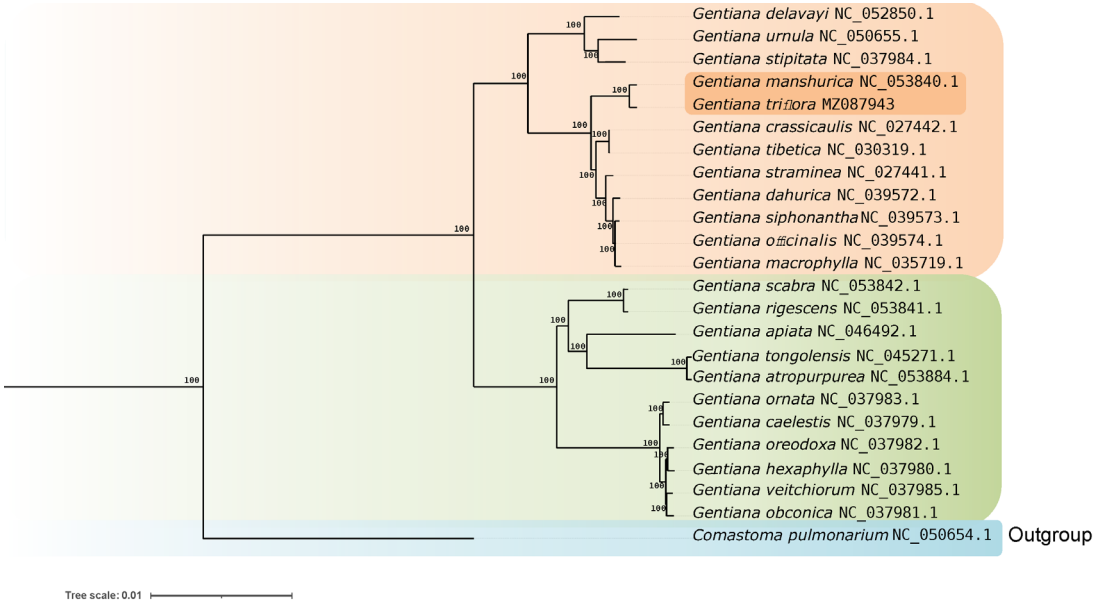


Fig. 7: Phylogenetic analysis of 23 species of *Gentiana* and Taxon *Comastoma pulmonarium* as outgroup based on chloroplast genome sequences by RAxML and bootstrap support values near the branch. The *Gentiana* genus was grouped into two clades (presented with different background colors). The *G. triflora* was closely related to the *G. manshurica*.

Discussion

The single circular cp genome structure of *G. triflora* was similar to other sequenced cp genomes of *Gentiana* species with a typical quadripartite structure. The GC content (37.61%) was unevenly distributed along the cp genome, of which IR regions was higher than that of LSC and SSC, possibly due to the presence of GC-rich rRNA genes and tRNA in IR regions. The gene content showed that the cp genome of *G. triflora* is highly conserved and has no gene loss. However, some previous studies showed that *ndh* gene loss occurs in the plastid genomes of some species in *Gentianaceae*. The plants in high altitudes (altitude ≥ 3000 m) and cold regions may result in deletion of *ndh* gene (Fu et al., 2021). By contrast, *G. triflora* grows at altitudes between 640 and 950 m, which means that its distribution is fairly widespread and not restricted to colder, higher altitudes. The *ndh* complex is more important for electron cycling around photosystem I under hot pressure conditions than under cold pressure conditions (Wang et al., 2006). This phenomenon suggests that the loss of the *ndh* gene in plants may be related to variation in growth environment and distribution range.

SSRs are short and tandem repeat and distribute widely throughout the cp genome, which has been often used as molecular markers for species authentication due to high rates of variability (Provan, 2010). The SSRs analysis revealed that the majority of SSRs in the cp genomes are mononucleotide, the largest number was found in *G. triflora* of which most are poly T and A as in most flowering plants cp genome (Li et al., 2019). Most of the SSRs were found in non-coding genes regions whereas few were located in the protein-coding genes regions. The SSRs detected in this study will be a valuable resource that can be useful for the study of genetic diversity, molecular phylogeny and evolution in other related species (Su et al., 2015).

The contraction and expansion of IR regions borders are considered to be the main reasons for cp genome length changes (Goulding et al., 1996). Furthermore, with the expansion or contraction of the IR boundary regions, genes near the border have the opportunity to access IR or SC regions. In general, the cp genome sizes of species from the same section or genus are extremely homogeneous. The results are shown in Fig. 4 support this conclusion (Fu et al., 2021). The comparison of IR boundary regions among *G. triflora* with those of closely related species showed there was no significant difference, except for position changes in *ycf*1 and *rps*19. Overall, *G. macrophylla*, *G. caelestis*, *G. ornate*, *G. obconica*, *G. oreodoxa*, *G. hexaphylla,* and *G. veitchiorum* showed a higher variability than other species in the *Gentiana* genus.

The cp genomes of *Gentiana* genus showed less variation in non-coding genes regions than in the coding regions. The IR regions were least divergent, possibly due to the presence of highly conserved rRNA sequences in these regions. We recommend eight hypervariable regions, *trnH*-GUG-*psb*A, *trnG*-UCC-intron, *atp*I, *trnD*-GUC, *trnL*-UAA, *rpl*32-*trnL*-UAG, *pet*A, and *ycf*1, as potential molecular markers of the *Gentiana* genus. Most of the variable regions are located in the single-copy region particularly the LSC region, which is consistent in most angiosperms.

The complete cp genome provides abundant resources for inferring evolutionary and phylogenetic relationships (Dong et al., 2012; Borsch and Quandt, 2009; Tong et al., 2016). The phylogenetic relationship of *Gentiana* genus has long been debated. In this study, we performed phylogenetic analyses among *Gentiana* genus using complete cp genomes to estimate taxonomic and evolutionary relationships. Notably, contrary to previous phylogenetic analyses based on only a few genes, *G. triflora* had the closest relationship with *G. manshurica* and belonged to sect. *Pneumonanthe*. The phylogenetic tree had high bootstrap support in all nodes, showing the accuracy and the reliability of the phylogenetic relationships. Our results based on the complete cp genome, are expected to resolve the taxonomic status and evolutionary relationships of *Gentiana* genus in the future.

Conclusion

In our study, we sequenced and annotated the complete cp genome of *G. triflora*, providing valuable resources for further research in this species. Overall, the characterization of *G. triflora* cp genome was similar to other species of *Gentiana* genus. The SSRs resource developed in this study will be used to study the identification and evolution relationships. Moreover, 8 highly variable regions were detected, which provided potential markers for future barcoding and phylogenetic studies of *Gentiana*. The results of this study provide valuable resources for the study of the cp genome of *G. triflora,* which have important implications for investigating the *Gentiana* genus evolution and diversification.

Authors' contributions

LFH and YJM conceived the study and designed experiments; LFH collected the samples; XKZ assembled and annotated the cp genomes; YJ, TXZ and YZ carried out the comparative chloroplast analysis; XKZ analyzed the data and prepared a draft of the manuscript and figures. XS, and XKZ modified the manuscript. All authors have read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (U1812403-1 and 82073960), Chengdu University of Traditional Chinese Medicine (003109034001), Beijing Natural Scientific Foundation (7202135) and National Science & Technology Fundamental Resources Investigation Program of China (2018FY100701).

Availability of data and materials

The dataset generated and or analyzed during the current study is deposited in the genebank with accession number: MZ087943. The phylogenetic genome datasets used and analyzed in this study were retrieved from the National Center for Biotechnology Information Organelle Genome Resource Database.

Conflicts of interests

No potential conflict of interest was reported by the authors.

References

Aldridge, C., Maple, J., Møller, S.G., 2005: The molecular biology of plastid division in higher plants. J. Exp. Bot. 56(414), 1061-1077. DOI: 10.1093/jxb/eri118

Allen, G.C., Flores-Vergara, M., Krasynanski, S., Kumar, S., Thompson, W., 2006: A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. Nature Prot. 1(5), 2320-2325. DOI: 10.1038/nprot.2006.384

Amiryousefi, A., Hyvönen, J., Poczai, P., 2018: IRscope: an online program to visualize the junction sites of chloroplast genomes. Bioinformatics, 34(17), 3030-3031. DOI: 10.1093/bioinformatics/bty220

Beier, S., Thiel, T., Münch, T., Scholz, U., Mascher, M., 2017: MISA-web: a web server for microsatellite prediction. Bioinformatics 33(16), 2583-2585. DOI: 10.1093/bioinformatics/btx198

Borsch, T., Quandt, D., 2009: Mutational dynamics and phylogenetic utility of noncoding chloroplast DNA. Plant System Evol. 282(3-4), 169-199. DOI: 10.1371/journal.pone.0157183

Boudreau, E., Turmel, M., 1995: Gene rearrangements in Chlamydomonas chloroplast DNAs are accounted for by inversions and by the expansion/contraction of the inverted repeat. Plant Mol. Biol. 27(2), 351-364. DOI: 10.1007/BF00020189

Dierckxsens, N., Mardulyn, P., Smits, G., 2017: NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucl. Acids Res. 45(4), e18-e18. DOI: 10.1093/nar/gkw955

Dong, W., Liu, J., Yu, J., Wang, L., Zhou, S., 2012: Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. PloS one 7(4), e35071. DOI: 10.1371/journal.pone.0035071

Doi, H., Takahashi, R., Hikage, T., Takahata, Y., 2010: Embryogenesis and doubled haploid production from anther culture in gentian (*Gentiana triflora*). Plant Cell Tissue Organ Culture 102(1), 27-33. DOI: 10.1007/s11240-010-9700-1

Dubchak, I., Ryaboy, D.V., 2005: VISTA family of computational tools for comparative analysis of DNA sequences and whole genomes. Methods Mol. Biol. 338, 69-89. DOI: 10.1385/1-59745-097-9:69

Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M., Dubchak, I., 2004: VISTA: computational tools for comparative genomics. Nucl. Acids Res. 32(suppl_2), 273-279. DOI: 10.1093/nar/gkh458

Fu, P.C., Sun, S.S., Twyford, A.D., Li, B.B., Zhou, R.Q., Chen, S.L., Gao, Q.B., Favre, A., 2021: Lineage-specific plastid degradation in subtribe Gentianinae (Gentianaceae). Ecol. Evol. 11(7), 3286-3299. DOI: 10.1002/ece3.7281

Goulding, S.E., Wolfe, K.H., Olmstead, R.G., Morden, C.W., 1996: Ebb and flow of the chloroplast inverted repeat. Mol. General Genetics 252(1), 195-206. DOI: 10.1007/BF02173220

Greiner, S., Lehwark, P., Bock, R., 2019: OrganellarGenomeDRAW (OGDRAW) version 1.3. 1: expanded toolkit for the graphical visualization of organellar genomes. Nucl. Acids Res. 47(W1), 59-64. DOI: 10.1093/nar/gkz238

Jin, J.-J., Yu, W.-B., Yang, J.-B., Song, Y., DePamphilis, C.W., Yi, T.-S., Li, D.-Z., 2020: GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol. 21(1), 1-31. DOI: 10.1186/s13059-020-02154-5

John, R., Li, S., Mar, A.K., Standley, D.M., Kazutaka, K., 2019: MAFFT-DASH: integrated protein sequence and structural alignment. Nucl. Acids Res. (W1), 5-10. DOI: 10.1093/nar/gkz342

Kurtz, S., Choudhuri, J.V., Ohlenbusch, E., Schleiermacher, C., Stove, J., Giegerich, R., 2001: REPuter: the manifold applications of repeat analysis on a genomic scale. Nucl. Acids Res. 22, 4633-4642. DOI: 10.1093/nar/29.22.4633

Lewis, S.E., Searle, S., Harris, N., Gibson, M., Iyer, V., Richter, J., Wiel, C., Bayraktaroglu, L., Birney, E., Crosby, M., 2002: Apollo: a sequence annotation editor. Genome BIOL 3(12), 1-14. DOI: 10.1186/gb-2002-3-12-research0082

Li, X., Tan, W., Sun, J., Du, J., Zheng, C., Tian, X., Zheng, M., Xiang, B., Wang, Y., 2019: Comparison of four complete chloroplast genomes of medicinal and ornamental Meconopsis species: genome organization and species discrimination. Scientific reports 9(1), 1-12. DOI: 10.1038/s41598-019-47008-8

Liang, Y., Meng, X., Mu, Z., Qian, J., Duan, B., Xu, L., 2020: The complete chloroplast genome and phylogenetic analysis of *Gentiana manshurica* Kitag from China. Mitochondrial DNA Part B 5(2), 1625-1626. DOI: 10.1080/23802359.2020.1745107

Mishiba, K.I., Yamane, K., Nakatsuka, T., Nakano, Y., Nishihara, M., 2009: Genetic relationships in the genus Gentiana based on chloroplast DNA sequence data and nuclear DNA content. Breeding Sci. 59(2), 119-127. DOI: 10.1270/jsbbs.59.119

Mower, J.P., 2005: PREP-Mt: predictive RNA editor for plant mitochondrial genes. BMC bioinformatics 6(1), 1-15. DOI: 10.1186/1471-2105-6-96

Mower, J.P., 2009: The PREP suite: predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. Nucl. Acids Res. 37(suppl_2), 253-259. DOI: 10.1093/nar/gkp337

Neuhaus, H.E., Emes, M.J., 2000: Nonphotosynthetic metabolism in plastids. Annu. Rev. Plant Physiol. Plant Mol. Biol. 51(51), 111-140. DOI: 10.1146/annurev.arplant.51.1.111

Palmer, J.D., 1991: CHAPTER 2 – Plastid Chromosomes: Structure and Evolution. Molecular Biology of Plastids 1997, 5-53.

Provan, J., 2010: Novel chloroplast microsatellites reveal cytoplasmic variation in *Arabidopsis thaliana*. Mol. Ecol. 9(12), 2183-2185.

Rozas, J., 2009: DNA sequence polymorphism analysis using DnaSP. In: Bioinformatics for DNA sequence analysis, 337-350. Springer.

Shang, M., Meng, X., Yan, F., Qian, J., Duan, B., Wang, Y., 2020: Assembly and phylogenetic analysis of the complete chloroplast genome sequence of *Gentiana scabra* Bunge. Mitochondrial DNA Part B 5(2), 1691-1692. DOI: 10.1080/23802359.2020.1748538

Sharp, P.M., Li, W.-H., 1986: Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. Nucl. Acids Res. 14(19), 7737-7749. DOI: 10.1093/nar/14.19.7737

Sharp, P.M., Li, W.H., 1987: The codon Adaptation Index – a measure of directional synonymous codon usage bias, and its potential applications. Nucl. Acids Res. 15(3), 1281-1295. DOI: 10.1093/nar/15.3.1281

Shi, L., Chen, H., Jiang, M., Wang, L., Wu, X., Huang, L., Liu, C., 2019: CPGAVAS2, an integrated plastome sequence annotator and analyzer. Nucl. Acids Res. 47(W1), 65-73. DOI: 10.1093/nar/gkz345

Stamatakis, A., 2014: RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9), 1312-1313. DOI: 10.1093/bioinformatics/btu033

Song, Q., Gao, K., Fu, K., 1987: Isolation and identification of gentiopicroside from the roots of Gentiana triflora Pall. Zhong yao Tong bao (Beijing, China: 1981) 12(12), 36-37, 59.

Su, Y., Liu, Y., Li, Z., Fang, Z., Yang, L., Zhuang, M., Zhang, Y., 2015: QTL analysis of head splitting resistance in cabbage (*Brassica oleracea* L. var. capitata) using SSR and InDel makers based on whole-genome re-sequencing. Plos One 10(9), e0138073. DOI: 10.1371/journal.pone.0138073

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013: MEGA6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30(12), 2725-2729. DOI: 10.1093/molbev/mst197

Tong, W., Kim, T.-S., Park, Y.-J., 2016: Rice chloroplast genome variation architecture and phylogenetic dissection in diverse Oryza species assessed by whole-genome resequencing. Rice 9(1), 1-13. DOI: 10.1186/s12284-016-0129-y

Wang, C.-H., Wang, Z.-T., Annie Bligh, W., White, K.N., White, C.J.B., 2004: Pharmacokinetics and tissue distribution of gentiopicroside following oral and intravenous administration in mice. Eur. J. Drug Metabol. Pharmacokinet. 29(3), 199-203. DOI: 10.1007/BF03190598

Wang, P., Duan, W., Takabayashi, A., Endo, T., Shikanai, T., Ye, J.-Y., Mi, H., 2006: Chloroplastic NAD (P) H dehydrogenase in tobacco leaves functions in alleviation of oxidative damage caused by temperature stress. Plant Physiol. 141(2), 465-474. DOI: 10.1104/pp.105.070490

Wicke, S., Schneeweiss, G.M., De Pamphilis, C.W., Kai, F.M., Quandt, D., 2011: The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol. Biol. 76(3-5), 273-297. DOI: 10.1007/s11103-011-9762-4

Yamada, H., Kikuchi, S., Inui, T., Takahashi, H., Kimura, K.I., 2014: Gentiolactone, a Secoiridoid Dilactone from Gentiana triflora, Inhibits TNF-α, iNOS and Cox-2 mRNA Expression and Blocks NF-κB Promoter Activity in Murine Macrophages. Plos One 9(11), e113834. DOI: 10.1371/journal.pone.0113834

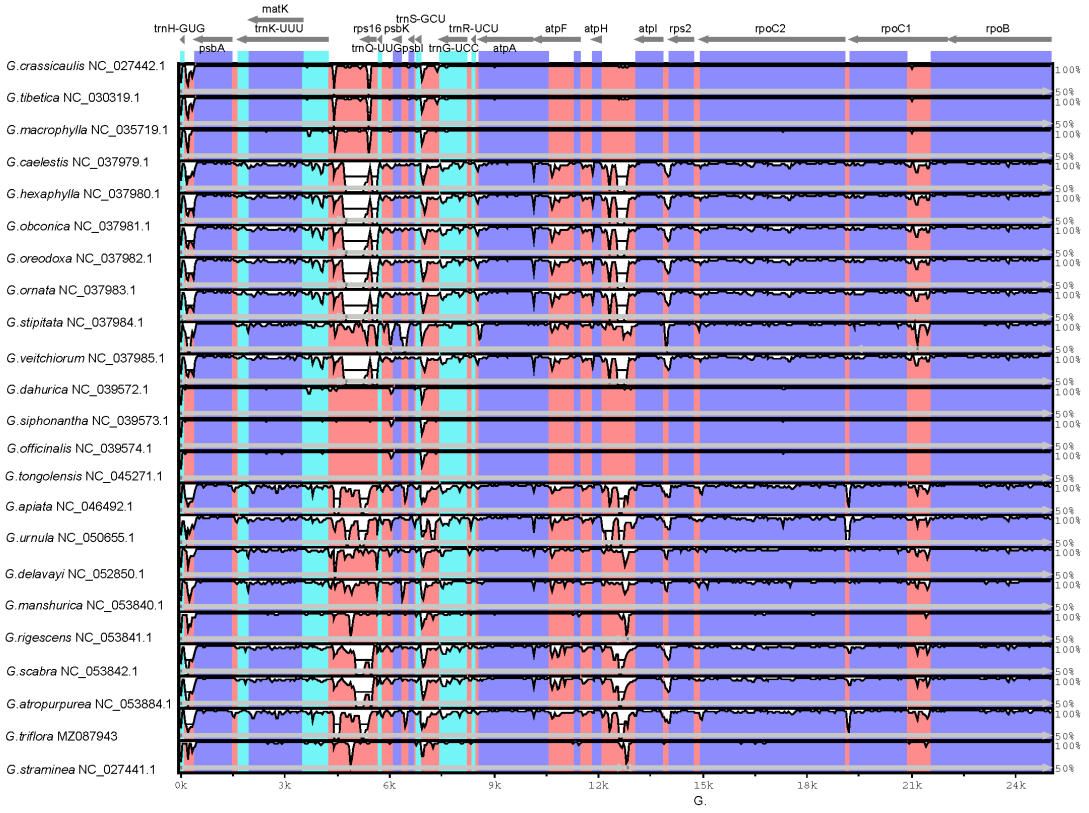
Yang, Q., Qian, J., Wang, J., Du, Z., Duan, B., 2020: The first complete chloroplast genome of *Gentiana rigescens* and its phylogenetic position in Gentianaceae. Mitochondrial DNA Part B 5(2), 1603-1604. DOI: 10.1080/23802359.2020.1745102

Yang, Z., 2007: PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. 24(8), 1586-1591. DOI: 10.1093/molbev/msm088

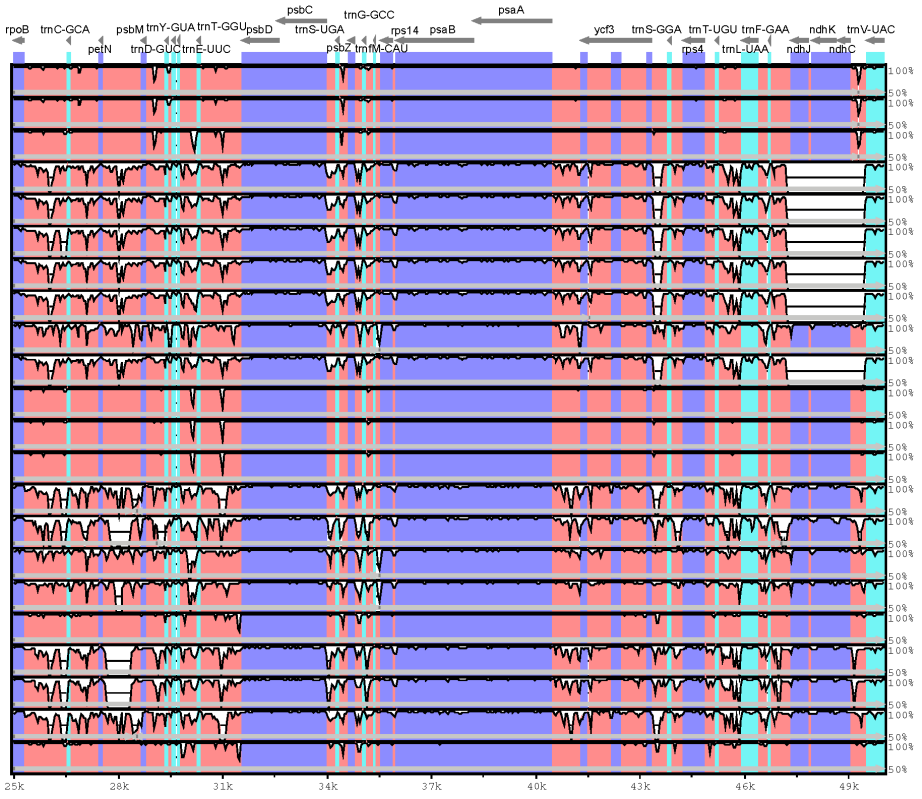
Yuan, Y.M., Küpfer, P., Doyle, J.J., 1996: Infrageneric phylogeny of the genus Gentiana (Gentianaceae) inferred from nucleotide sequences of the internal transcribed spacers (ITS) of nuclear ribosomal DNA. Am. J. Bot. 83(5), 641-652. DOI: 10.1002/j.1537-2197.1996.tb12750.x

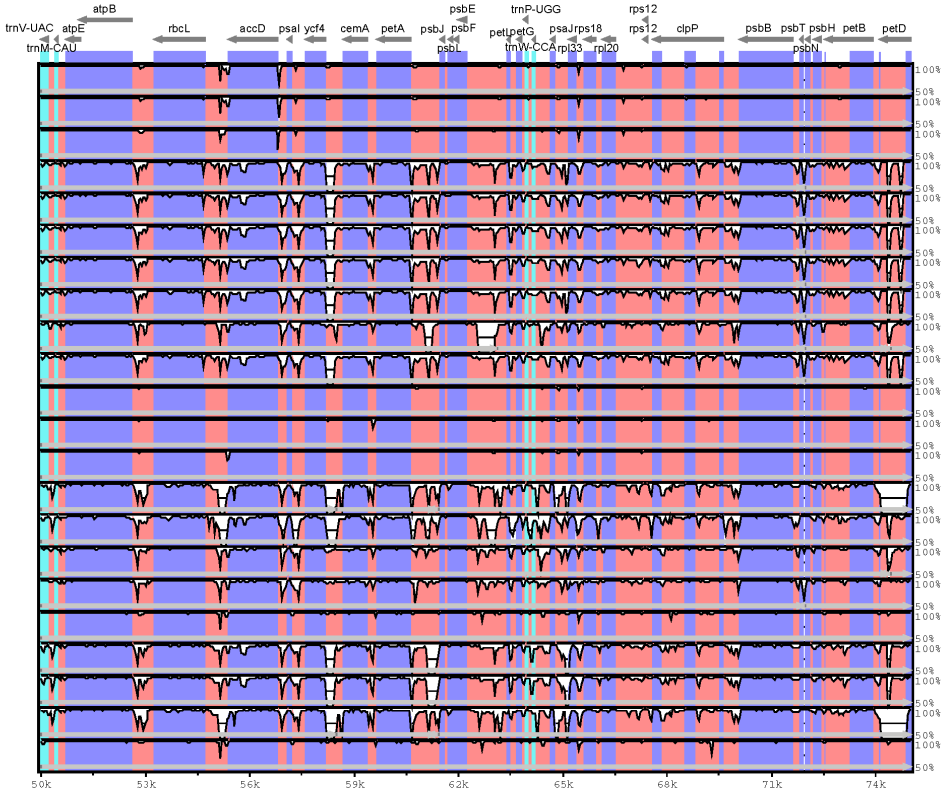
Zhang, D., Gao, F., Jakovlić, I., Zou, H., Zhang, J., Li, W.X., Wang, G.T., 2020: PhyloSuite: An integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. Mol. Ecol. Res. 20(1), 348-355. DOI: 10.1111/1755-0998.13096

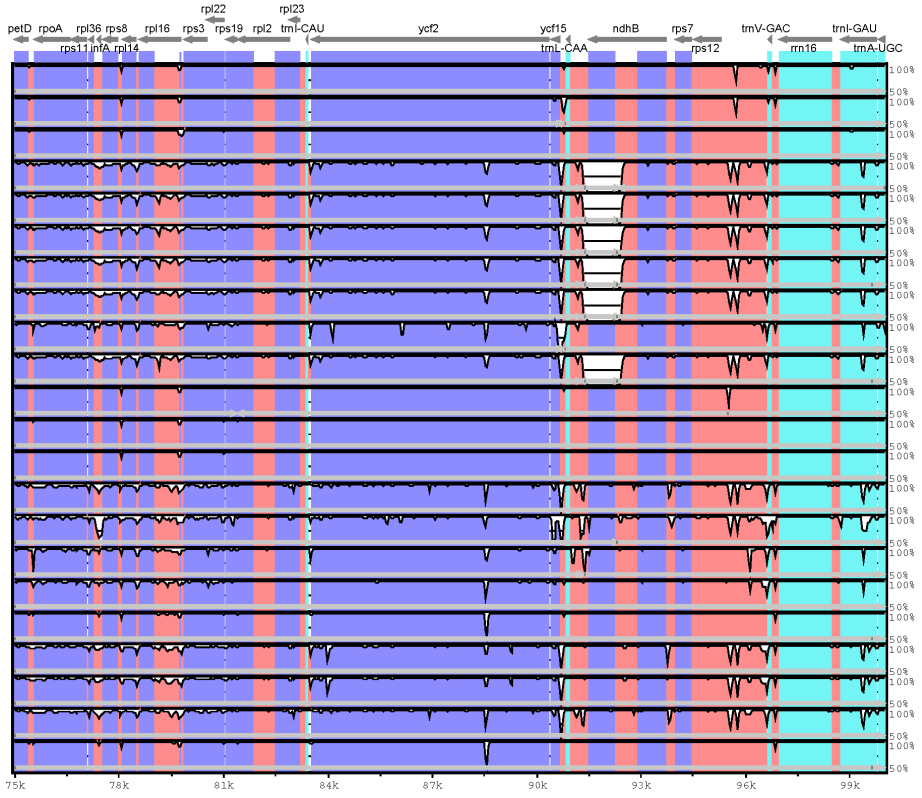
Address of the corresponding author:
LinFang Huang, Key lab of Chinese Medicine Resources Conservation, State Administration of Traditional Chinese Medicine of China, Institute of Medicinal Plant Development, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing 100193, China
E-mail: lfhuang@implad.edu.cn

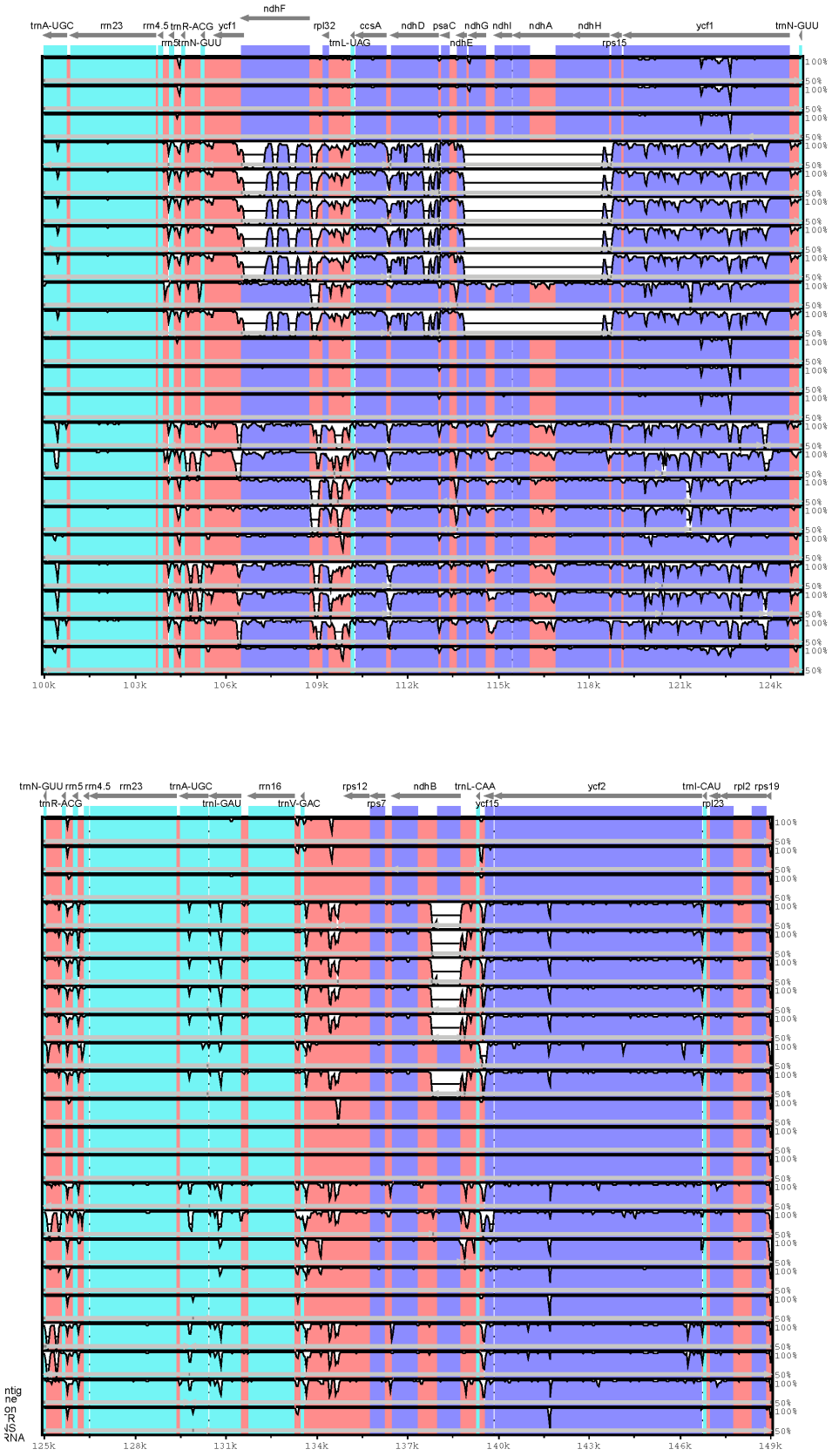

matK
trnH-GUG
trnK-UUU
rps16
psbK
trnS-GCU
trnR-UCU
atpF
atpH
atpI
rps2
rpoC2
rpoC1
rpoB
psbA
trnQ-UUG
psbI
trnG-UCC
atpA
G.crassicaulis NC_027442.1
G.tibetica NC_030319.1
G.macrophylla NC_035719.1
G.caelestis NC_037979.1
G.hexaphylla NC_037980.1
G.obconica NC_037981.1
G.oreodoxa NC_037982.1
G.ornata NC_037983.1
G.stipitata NC_037984.1
G.veitchiorum NC_037985.1
G.dahurica NC_039572.1
G.siphonantha NC_039573.1
G.officinalis NC_039574.1
G.tongolensis NC_045271.1
G.apiata NC_046492.1
G.urnula NC_050655.1
G.delavayi NC_052850.1
G.manshurica NC_053840.1
G.rigescens NC_053841.1
G.scabra NC_053842.1
G.atropurpurea NC_053884.1
G.triflora MZ087943
G.straminea NC_027441.1
100%
50%
0k
3k
6k
9k
12k
15k
18k
21k
24k

G.


psbC
trnG-GCC
psaA
rpoB
trnC-GCA
psbM
trnY-GUA
trnT-GGU
psbD
trnS-UGA
rps14
psaB
ycf3
trnS-GGA
trnT-UGU
trnF-GAA
ndhK
trnV-UAC
petN
trnD-GUC
trnE-UUC
psbZ
trnfM-CAU
rps4
trnL-UAA
ndhJ
ndhC
100%
50%
25k
28k
31k
34k
37k
40k
43k
46k
49k


atpB
trnV-UAC
atpE
trnM-CAU
rbcL
accD
psaI
ycf4
cemA
petA
psbJ
psbF
psbL
psbE
petL
petG
trnP-UGG
trnW-CCA
psaJ
rps18
rpl33
rpl20
rps12
rps12
clpP
psbB
psbT
psbN
psbH
petB
petD
100%
50%
50k
53k
56k
59k
62k
65k
68k
71k
74k


petD
rpoA
rps11
rpl36
infA
rps8
rpl14
rpl16
rps3
rpl22
rps19
rpl2
rpl23
trnI-CAU
ycf2
ycf15
trnL-CAA
ndhB
rps7
rps12
trnV-GAC
rrn16
trnI-GAU
trnA-UGC
100%
50%
75k
78k
81k
84k
87k
90k
93k
96k
99k



Supplement Fig. 1: Comparison of the plastomes in Gentiana genus by using mVISTA. The genes are represented as gray arrows on the top of the alignments. The different regions are labeled with different colors. The pink regions are "conserved noncoding sequences" (CNS), the dark blue regions are exons, and the light-blue regions are tRNAs or rRNAs. The percentages 50 and 100% refer to the similarity among sequences. The gray arrows above the aligned sequences represent genes and their orientation

Supplement Table 1 Predicted RNA editing site in the *G. triflora* chloroplast genome.

Gene	Nt Pos	AA Pos	Align Col	Effect	Score
accD (C = 0.8)	175	59	60	CCT (P) => TCT (S)	1
	241	81	88	CTT (L) => TTT (F)	1
atpA (C = 0.8)	773	258	258	TCA (S) => TTA (L)	1
	791	264	264	CCC (P) => CTC (L)	1
atpI (C = 0.8)	626	209	213	TCG (S) => TTG (L)	1
matK (C = 0.8)	25	109	125	CCT (P) => TCT (S)	1
ccsA (C = 0.8)	445	149	165	CAT (H) => TAT (Y)	1
	872	291	307	CCT (P) => CTT (L)	0.86
	1243	415	431	CAT (H) => TAT (Y)	1
	1493	498	515	ACT (T) => ATT (I)	1
ndhA (C = 0.8)	350	117	117	TCA (S) => TTA (L)	1
	575	192	192	TCA (S) => TTA (L)	1
ndhB (C = 0.8)	149	50	50	TCA (S) => TTA (L)	1
	467	156	156	CCA (P) => CTA (L)	1
	586	196	196	CAT (H) => TAT (Y)	1
	611	204	204	TCA (S) => TTA (L)	0.8
	737	246	246	CCA (P) => CTA (L)	1
	746	249	249	TCT (S) => TTT (F)	1
	830	277	277	TCA (S) => TTA (L)	1
	836	279	279	TCA (S) => TTA (L)	1
	1481	494	494	CCA (P) => CTA (L)	1
ndhD (C = 0.8)	44	15	15	ACG (T) => ATG (M)	1
	89	30	30	TCT (S) => TTT (F)	0.8
	425	142	142	TCA (S) => TTA (L)	1
	716	239	239	TCA (S) => TTA (L)	1
	868	290	290	CTT (L) => TTT (F)	1
	920	307	307	TCA (S) => TTA (L)	1
	1301	434	434	TCT (S) => TTT (F)	1
	1340	447	447	TCA (S) => TTA (L)	0.8
	1352	451	451	TCA (S) => TTA (L)	0.8
ndhF (C = 0.8)	290	97	97	TCA (S) => TTA (L)	1
	1712	571	584	TCG (S) => TTG (L)	0.8
	2215	739	752	CCC (P) => TTC (F)	1
	2216	739	752	CCC (P) => TTC (F)	1
ndhG (C = 0.8)	314	105	105	ACA (T) => ATA (I)	0.8
petB (C = 0.8)	611	204	204	CCA (P) => CTA (L)	1
psaI (C = 0.8)	80	27	43	TCT (S) => TTT (F)	0.86
rpl20 (C = 0.8)	308	103	103	TCA (S) => TTA (L)	0.86
rpoB (C = 0.8)	157	53	53	CTT (L) => TTT (F)	1
	473	158	159	TCA (S) => TTA (L)	0.86
	551	184	185	TCA (S) => TTA (L)	1
	566	189	190	TCG (S) => TTG (L)	1
	2432	811	829	TCA (S) => TTA (L)	0.86
rpoC1 (C = 0.8)	41	14	14	TCA (S) => TTA (L)	1
	331	111	111	CGC (R) => TGC (C)	1
rpoC2 (C = 0.8)	730	244	255	CTT (L) => TTT (F)	0.86
	1294	432	449	CCT (P) => TCT (S)	0.86
	1921	641	841	CAC (H) => TAC (Y)	1
	2248	750	952	CGG (R) => TGG (W)	1
	2689	897	1106	CCC (P) => TCC (S)	1
	1493	498	515	ACT (T) => ATT (I)	1
	2773	925	1134	CCA (P) => TCA (S)	1
	2819	940	1149	TCC (S) => TTC (F)	0.86
	3017	1006	1223	GCA (A) => GTA (V)	0.86
	3683	1228	1455	TCA (S) => TTA (L)	0.86
rps2 (C = 0.8)	248	83	83	TCA (S) => TTA (L)	1
rps14 (C = 0.8)	149	50	53	CCA (P) => CTA (L)	1