

Article

Not peer-reviewed version

Complete chloroplast genomes of *Microula sikkimensis* and comparative analyses with related species from Boraginaceae

[Yunqing Gao](#) , [Zhenjiang Chen](#) ^{*} , [Xiuzhang Li](#) , [Kamran Malik](#) , [Chunjie Li](#) ^{*}

Posted Date: 28 December 2023

doi: 10.20944/preprints202312.2224.v1

Keywords: *Microula sikkimensis*; chloroplast genomes; comparative analysis; phylogenetic analysis; Boraginaceae



Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Article

Complete Chloroplast Genomes of *Microula sikkimensis* and Comparative Analyses with Related Species from Boraginaceae

Yunqing Gao ^{1,2,3}, Zhenjiang Chen ^{1,2,3,*}, Xiuzhang Li ⁴, Kamran Malik ^{1,2,3} and Chunjie Li ^{1,2,3,*}

¹ State Key Laboratory of Herbage Improvement and Grassland Agro-Ecosystems, Lanzhou University, Lanzhou 730020, China; 220220902301@lzu.edu.cn (Y.G.); malik@lzu.edu.cn (K.M.)

² Key Laboratory of Grassland Livestock Industry Innovation, Ministry of Agriculture and Rural Affairs, Lanzhou University, Lanzhou 730020, China

³ College of Pastoral Agricultural Science and Technology, Lanzhou University, Lanzhou 730020, China

⁴ Qinghai Academy of Animal and Veterinary Science, Qinghai University, Xining 810016, China; xiuzhang11@163.com

* Correspondence: chenzhenjiang@lzu.edu.cn (Z.C.); chunjie@lzu.edu.cn (C.L.)

Abstract: The present study provides a detailed analysis of the chloroplast genome of *Microula sikkimensis*. The genome was consisted of altogether 149,428bp and distinct regions, including large single-copy (81329bp), small single-copy (17261bp), and inverted repeat (25419bp). The genome contained 112 genes, including 78 protein-coding genes, 30 tRNA genes, and 4 rRNA genes, some exhibited duplication in the IR region. The chloroplast genome displayed different GC content across regions, with the IR region exhibiting the highest. Codon usage analysis and the identification of simple sequence repeats (SSRs) offer valuable genetic markers. Comparative analysis with other Boraginaceae species highlighted conservation and diversity in coding and noncoding regions. Phylogenetic analysis placed *M. sikkimensis* within the Boraginaceae family, revealing its distinct relationship with specific species.

Keywords: *Microula sikkimensis*; chloroplast genomes; comparative analysis; phylogenetic analysis; Boraginaceae

1. Introduction

M. sikkimensis, a biennial herbaceous plant belonging to the family Boraginaceae, is predominantly found in the high-altitude grasslands, forests, shrublands, and secondary vegetation at elevations ranging from 2500 to 4000 meters on the eastern edge of the Qinghai-Tibet Plateau [1]. *M. sikkimensis* is a plant resource rich in γ -linolenic acid, with promising development prospects. Abundant experiments have confirmed that *M. sikkimensis* oil significantly reduces the levels of total cholesterol (TC), triglycerides (TG), and serum malondialdehyde (MDA) in the liver and serum. It also increases the ratio of high-density lipoprotein cholesterol to total cholesterol (HDL-C/TC) in the serum. Moreover, it effectively reduces the deposition of cholesterol in peripheral tissue cells, thus preventing atherosclerosis and maintaining the integrity of the biomembrane structure. Its capacity to lower triglycerides in the liver and serum surpasses that of atorvastatin. *M. sikkimensis* oil can improve the high blood lipidemia by reducing blood viscosity, preventing thrombosis, and exhibiting unique solvent properties [1–3]. The harvested stalks of *M. sikkimensis* exhibit heightened palatability and are nutritionally dense, making them a crucial coarse fodder for supplementing livestock in the winter and spring seasons in high-altitude pastoral regions. The research of *M. sikkimensis* oil will provide enough supply of raw materials for the production of a series of high-nutrition health foods, fortified dairy products, healthful edible oils, as well as specialized pharmaceuticals and novel cosmetic [4].

Chloroplasts are specialized energy converters unique to higher plants and certain algae, and they serve as essential organelles with autonomous genetic information within cells [5–7]. The plant chloroplast genome is generally a double-stranded circular molecule composed of four main regions: a large single copy region (LSC), a small single copy region (SSC), and two inverted repeat regions (IRA and IRB), with the sequences of the two IR regions being identical but in opposite directions [8–11]. The relatively small (115–165kb) and conservative nature of the chloroplast genome makes it a valuable tool for investigating genomic evolution and phylogenetic relationships in angiosperms [12–14]. Additionally, the chloroplast genome finds widespread applications in diverse domains, including population genetics, molecular-assisted breeding, gene mapping, plant barcode sequence screening and gene diversity studies [15–19].

Currently, there have been no research findings on the chloroplast genome of *M. sikkimensis*. In this study, we presented the first report and analysis of the complete chloroplast genome sequence of *M. sikkimensis*, by obtaining information on the basic genome structure, codon usage bias, and simple sequence repeats (SSRs). Additionally, comparative genomic analysis and phylogenetic analysis of *M. sikkimensis* chloroplast genome were conducted in relation to other species within the Boraginaceae family. These research findings could provide essential genetic references for the development of molecular markers based on the chloroplast genome and for the inference of phylogenetic relationships.

2. Materials and methods

2.1. Sample collection and DNA extraction

The *M. sikkimensis* plants were collected from Zaduo county, Qinghai province, China. The fresh leaves were stored in liquid nitrogen. Modified cetyltrimethylammonium bromide (CTAB) method was used to extract total genomic DNA. The quality of DNA was measured by NanoDrop 2000 (Thermo Scientific, Wilmington, NC, USA) and agarose gel electrophoresis.

2.2. Genome annotation

CPGAVAS2 software [20] was used to annotate the chloroplast genome while OGDRAW software [21] was used to visualize the circular chloroplast genome map of *M. sikkimensis*. The tRNA of the chloroplast genome was annotated using tRNAscanSE software (v.2.0.11) [22]. The rRNA of the chloroplast genome was annotated using BLASTN software (v2.13.0) [23]. The annotation errors of each chloroplast genome were manually corrected using CPGView software [24] and Apollo software (v1.11.8) [25]. The fully annotated chloroplast genome was finally deposited at the GenBank database (Accession Number: OR866440).

2.3. Comparative genome analysis

The complete chloroplast genomes of *M. sikkimensis* and other four species were compared using the MVISTA program [26] with the shuffle-LAGAN model, with *M. sikkimensis* as the reference. IRSCOPE [27] was applied to analyze the LSC, IR and LSC boundary locations in five Boraginaceae species complete chloroplast genomes.

The complete chloroplast genomes of the five Boraginaceae plants were multiple aligned by MAFFT [28], and nucleotide variations (Pi) was calculated by DnaSP [29] with the following parameters: window length, 600; step size, 200. The protein-coding sequences of genomes are extracted using Phylosuite software (v1.1.16) [30]. The codon preference of protein-coding genes in chloroplast genome was analyzed and the RSCU value was calculated using Mega software (v7.0) [31].

2.4. Repeat sequence analysis

Using MISA (v2.1) (<https://webblast.ipk-gatersleben.de/misa/>) [32], TRF(v4.09) (<https://tandem.bu.edu/trf/trf.unix.help.html>) [33] and REPuter web server (<https://bibiserv.cebitec.uni-bielefeld.de/reputer/>) [34] to identify the repeat sequences including

microsatellite sequence repeats, tandem repeats and scattered repeats. The results were visualized using Excel (2021) software.

2.5. Phylogenetic analysis

The complete chloroplast genomes of 30 species were downloaded from the National Center for Biotechnology Information (NCBI). *Isodon serra* and *Forsythia suspensa* were chosen as the outgroups. The detailed list of all the species along with their respective accession numbers for the chloroplast genomes available in the NCBI database could be found in Table S1. All the sequences were aligned by MAFFT with default parameter. Ambiguously aligned fragments were filtered by GBLOCKS [35], parameter settings: minimum number of sequences for a conserved position, 20; minimum number of sequences for a flank position, 20; maximum number of contiguous nonconserved positions, 6; minimum length of a block, 11; allowed gap positions, 0. Based on the result of alignment, phylogenetic tree was constructed using IQ-TREE [36] with 5000 bootstraps. Bayesian inference (BI) phylogenies were inferred using MRBAYES 3.2.0 [37] under the GTR+I+G model (eight parallel runs and 2 000 000 generations), in which the initial 25% of sampled data were discarded as burn-in. The generated trees were visualized using the online web tool iTOL [38].

3. Results

3.1. Chloroplast genome assembly and genome features

The total length of the chloroplast genome in *M. sikkimensis* is 149,428 bp, exhibiting a typical circular quadripartite structure composed of four distinct regions: a LSC region (81329bp), a SSC region (17261bp) and a pair of IR regions(25419bp of each) which separated SSC and LSC (Figure 1).

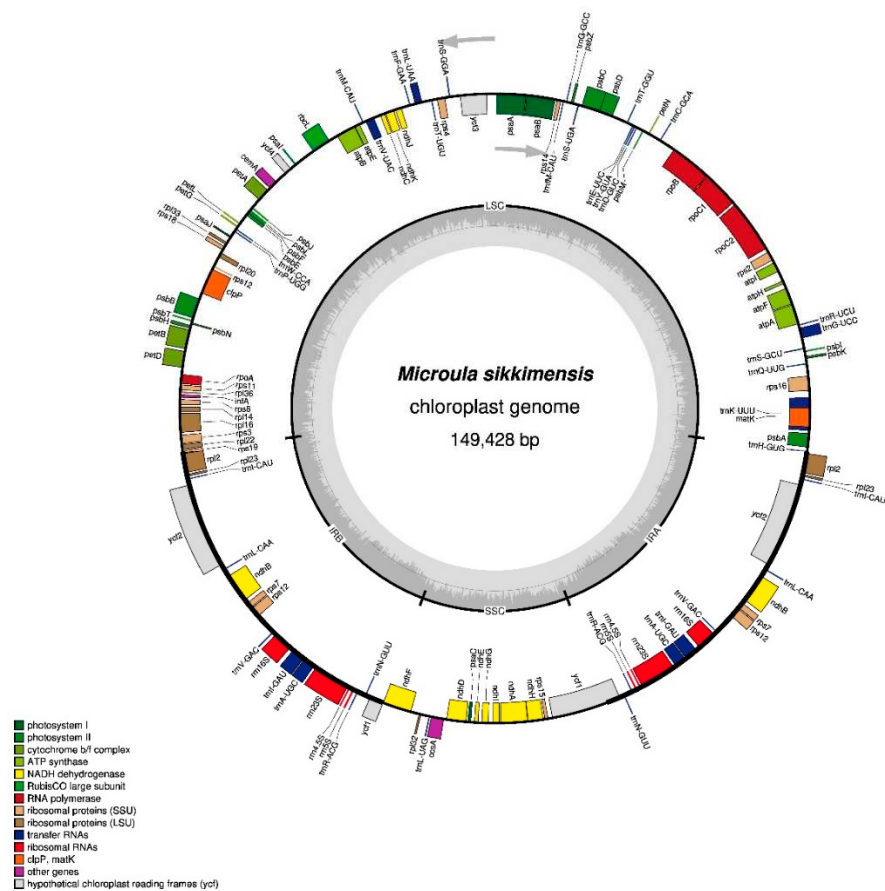


Figure 1. Chloroplast genome map of *M. sikkimensis*. Genes inside the circle are transcribed clockwise, while those outside the circle are transcribed anticlockwise. Large single copy (LSC) region, inverted repeat (IRA, IRB) regions and small single copy (SSC) region are shown in the figure. The darker gray

in the inner circle shows the GC content, while the lighter gray shows the AT content. Genes with different functions represented by different colors.

Gene annotation of the chloroplast genome in *M. sikkimensis* resulted in the identification of a total of 112 genes, including 78 protein-coding genes, 30 tRNA genes, and 4 rRNA genes. The protein-coding genes could be classified into 15 gene families: The protein coding genes included 15 gene families, including 11 NADH dehydrogenase genes (*ndhA*, *ndhB*, *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *ndhJ*, *ndhK*); 5 photosystem I genes (*psaA*, *psaB*, *psaC*, *psaI*, *psaJ*); 16 photosystem II genes (*psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbN*, *psbT*, *psbZ*, *ycf3*); 6 cytochrome b/f complex genes (*petA*, *petB*, *petD*, *petG*, *petL*, *petN*); 6 ATP synthase genes (*atpA*, *atpB*, *atpE*, *atpF*, *atpH*, *atpI*); 1 1, 5-diphosphate ribulose carboxylase/oxygenase large subunit gene (*rbcL*); 4 DNA-dependent RNA polymerase genes (*rpoA*, *rpoB*, *rpoC1*, *rpoC2*); 9 ribosome major genes (*rpl2*, *rpl14*, *rpl16*, *rpl20*, *rpl22*, *rpl23*, *rpl32*, *rpl33*, *rpl36*); 12 ribosome small genes (*rps2*, *rps3*, *rps4*, *rps7*, *rps8*, *rps11*, *rps12*, *rps14*, *rps15*, *rps16*, *rps18*, *rps19*); 1 mature enzyme gene (*matK*); 1 membrane protein gene (*cemA*); 1 protease gene (*clpP*); 1 C-type cytochrome synthesis gene (*ccsA*); 1 translation initiation factor (*infA*); 3 conserved open reading frame genes (*ycf1*, *ycf2*, *ycf4*) (Table 1).

Table 1. List of genes annotated in the chloroplast genomes of *M. sikkimensis*.

Group of genes	Name of genes	number of genes
NADH-dehydrogenase	<i>ndhA</i> , <i>ndhB</i> (×2), <i>ndhC</i> , <i>ndhD</i> , <i>ndhE</i> , <i>ndhF</i> , <i>ndhG</i> , <i>ndhH</i> , <i>ndhI</i> , <i>ndhJ</i> , <i>ndhK</i>	11
photosystem I	<i>psaA</i> , <i>psaB</i> , <i>psaC</i> , <i>psaI</i> , <i>psaJ</i>	5
photosystem II	<i>psbA</i> , <i>psbB</i> , <i>psbC</i> , <i>psbD</i> , <i>psbE</i> , <i>psbF</i> , <i>psbH</i> , <i>psbI</i> , <i>psbJ</i> , <i>psbK</i> , <i>psbL</i> , <i>psbM</i> , <i>psbN</i> , <i>psbT</i> , <i>psbZ</i> , <i>ycf3</i>	16
cytochrome b/f complex	<i>petA</i> , <i>petB</i> , <i>petD</i> , <i>petG</i> , <i>petL</i> , <i>petN</i>	6
ATP synthase	<i>atpA</i> , <i>atpB</i> , <i>atpE</i> , <i>atpF</i> , <i>atpH</i> , <i>atpI</i>	6
Large subunit of rubisco	<i>rbcL</i>	1
Small subunit of ribosome	<i>rps2</i> , <i>rps3</i> , <i>rps4</i> , <i>rps7</i> (×2), <i>rps8</i> , <i>rps11</i> , <i>rps12</i> (×2), <i>rps14</i> , <i>rps15</i> , <i>rps16</i> , <i>rps18</i> , <i>rps19</i>	12
Large subunit of ribosome	<i>rpl2</i> (×2), <i>rpl14</i> , <i>rpl16</i> , <i>rpl20</i> , <i>rpl22</i> , <i>rpl23</i> (×2), <i>rpl32</i> , <i>rpl33</i> , <i>rpl36</i>	9
DNA dependent RNA polymerase	<i>rpoA</i> , <i>rpoB</i> , <i>rpoC1</i> , <i>rpoC2</i>	4
rRNA genes	<i>rrn4.5S</i> (×2), <i>rrn5S</i> (×2), <i>rrn16S</i> (×2), <i>rrn23S</i> (×2)	4
tRNA genes	30tRNAs	30
Maturase	<i>matK</i>	1
Envelope membrane protein	<i>cemA</i>	1
Protease	<i>clpP</i>	1
c-type cytochrome synthesis gene	<i>ccsA</i>	1
Translational initiation factor	<i>infA</i>	1
Genes of unknown functions Open Reading	<i>ycf1</i> (×2), <i>ycf2</i> (×2), <i>ycf4</i>	3

Among the 112 genes, seven protein-coding genes (*ndhB*, *rpl2*, *rpl23*, *rps7*, *rps12*, *ycf1*, *ycf2*), seven tRNA genes (*trnA*-UGC, *trnL*-CAA, *trnL*-CAU, *trnL*-GAU, *trnN*-GUU, *trnR*-ACG, *trnV*-GAC), and four rRNA genes (*rrn4.5S*, *rrn5S*, *rrn16S*, *rrn23S*) were duplicated in the IR regions (Figure 1).

The chloroplast genome of *M. sikkimensis* exhibited a GC content of 37.51%. The GC content exhibited significant variations in different regions within the chloroplast genome. The IR region consistently displayed the highest GC content (43.13%), followed by the LSC region with a GC content of 35.38%, and the SSC region with a GC content of 30.99%.

In the chloroplast genome of *M. sikkimensis*, 12 genes contained introns, with 10 genes (rps12, rps16, atpF, rpoC1, petB, petD, rpl16, rpl2, ndhB and ndhA) containing only one intron and the 2 genes (ycf3 and clpP) harboring two introns. Among them, the intron of rpoC1 gene was the largest (1617 bp), while the intron of ycf3 gene was the smallest (153 bp) (Table S2).

Among the protein-coding genes of *M. sikkimensis*, unlike the conventional ATG initiation codon, the initiation codon of rps19 gene with GTG, while ndhD gene started with ACG (Table S3), which were similar to the other species [39–43].

3.2. Codon usage bias

The codon usage bias were analyzed based on the 78 protein-coding genes in *M. sikkimensis* chloroplast genome. Codons with a relative synonymous codon usage (RSCU) greater than 1 were considered as preferential for amino acid usage [44]. The majority of genes displayed codon preference, with the exception of the start codon AUG and the tryptophan codon UGG, both having an RSCU value of 1. The leucine amino acid exhibited the highest preference for the UUA codon (RSCU = 2) and the lowest preference for the CUG codon (RSCU = 0.32) among the protein-coding genes in the chloroplast of *M. sikkimensis*. (Figure 2)

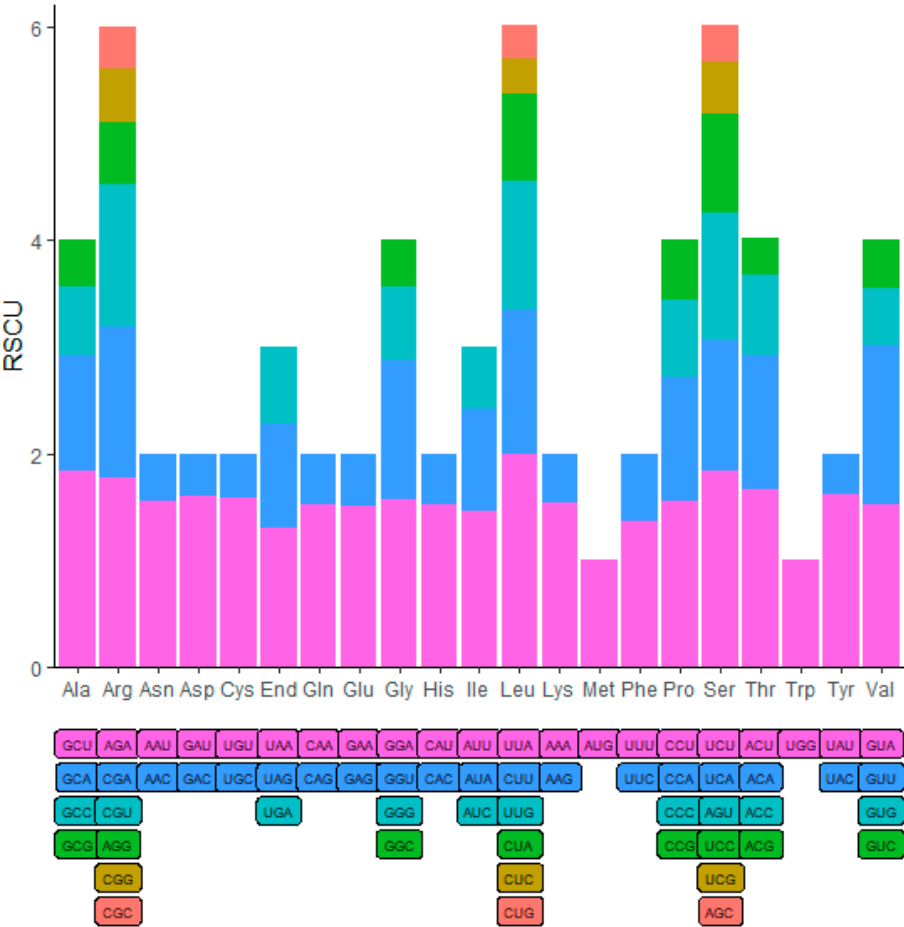


Figure 2. Codon content and RSCU value of the 20 amino acid and stop codons in all protein-coding genes of *M. sikkimensis* chloroplast genome.

3.3. Repeat sequences and SSR analysis

A total of 38 SSRs were identified in the chloroplast genome of *M. sikkimensis*, with one located in the IR region, 26 in the LSC region, and 11 in the SSC region. Mononucleotide repeats consisted of 10-12 repetitions, dinucleotide repeats had 5-8 repetitions, trinucleotide repeats had 4 repetitions, while tetra- and pentanucleotide repeats had 3 repetitions. SSRs composed of A/T motifs were more abundant than those composed of G/C motifs. Among the SSRs, A/T mononucleotide repeats were the most frequent (n = 22), followed by AT dinucleotide repeats (n = 8) (Table S4).

Tandem repeats, also known as satellite DNA, are core repeating units of about 7 to 200 bases that are repeated many times in tandem. Tandem repeats exist widely in eukaryotic genomes and prokaryotes. There are 18 tandem repeats in the *M. sikkimensis* chloroplast genome with a matching degree greater than 74% and a length between 8 and 24bp (Figure 3).

The dispersed repeat sequences in the chloroplast genome of *M. sikkimensis* were analyzed by REPuter. The results revealed a total of 36 pairs of repetitive sequences with a length equal to or greater than 30 bp. Among these pairs, 14 were identified as palindromic repeats, 18 as forward repeats, 3 as reverse repeats, and 1 as complementary repeat. The longest palindromic repeat was found to be 46 bp, while the longest forward repeat was 52 bp (Figure 3).

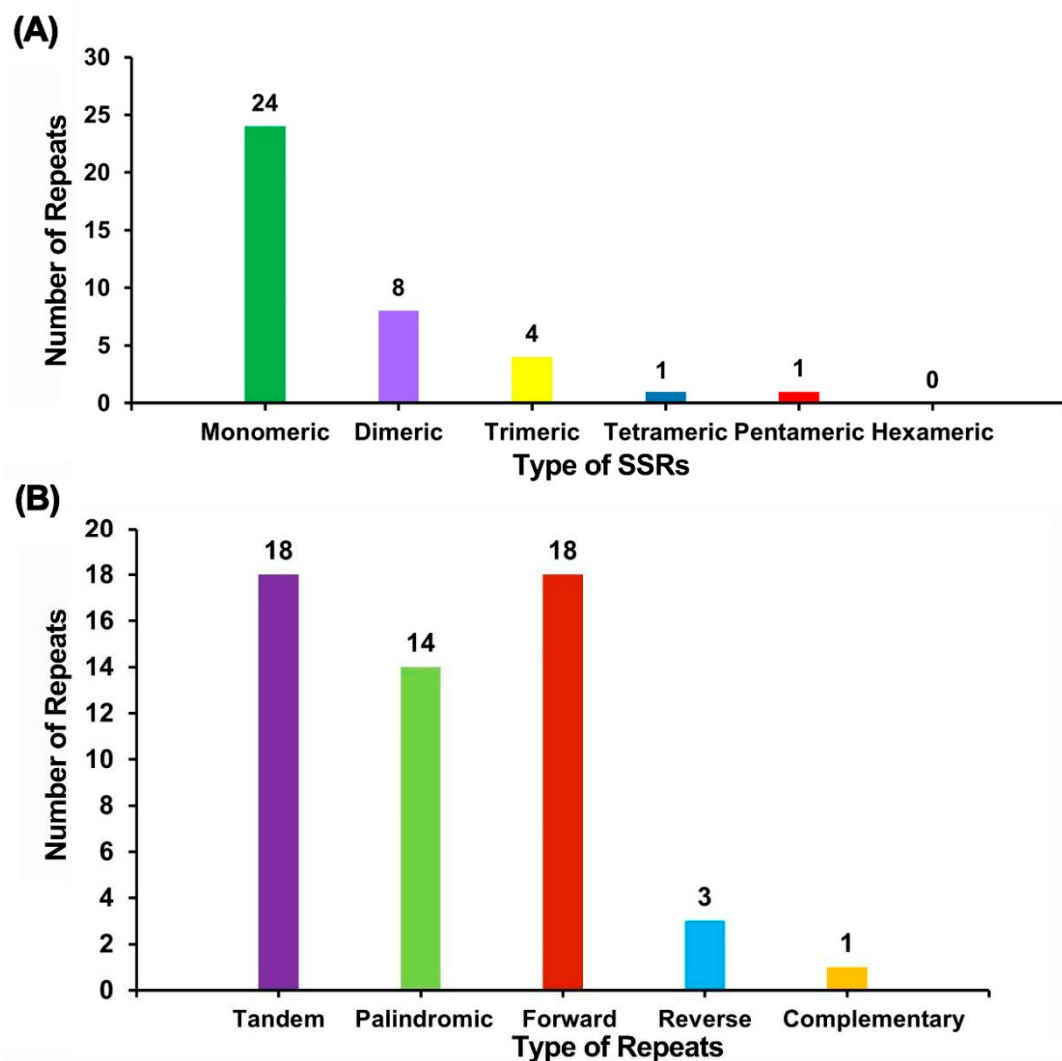


Figure 3. Repeat sequence and SSR analysis of *M. sikkimensis* chloroplast genome. (A) The horizontal coordinate represents the type of SSRs, the vertical coordinate represents the number of repeats, the green represents monomer SSRs, the purple represents dimer SSRs, the yellow represents trimer SSRs, the blue represents tetramer SSRs, and the red represents pentamer SSRs. No hexamer SSRs were detected in the chloroplast genome. (B) The horizontal coordinate indicates the type of repeat

sequence, the vertical coordinate indicates the number of repeat segments, purple indicates tandem repeats, green indicates palindromic repeats, red indicates forward repeats, blue indicates reverse repeats, and yellow indicates complementary repeats.

3.4. Comparative genome analysis

The chloroplast genomes of five Boraginaceae species were compared using mVISTA with *M.sikkimensis* as a reference. As a result, these five genomes were basically identical in coding regions, whereas more diverse in noncoding regions. The highly divergent regions were identified in intergenic spacers, including *matK-rps16*, *rps16-trnQ-UUG*, *trnS-GCU*, *trnF-GAA-ndhJ*, *rbcL-psaI*, *ycf4-cemA* and *petA-psbJ* in LSC, *ccsA-ndhD* and *rps15-ycf1* in SSC, which might be the molecular marks for Boraginaceae species (Figure 4).

To evaluate the degree of sequence variation, the chloroplast genome sequences of five Boraginaceae species were aligned, and the nucleotide variability (Pi) was calculated using DnaSP software. As shown in Figure 5, the Pi values ranged from 0 - 0.062, with an average value of 0.013. Four high Pi value (≥ 0.05) mutational hotspot loci were screened out in these five Boraginaceae species, including one protein coding gene (*ndhH*), one tRNA coding gene (*trnG-UCC*) and two intergenic regions (*trnQ-UUG-psbI* and *trnY-GUA-trnT-GUU*). These regions had the potential to be the molecular markers of Boraginaceae species. Furthermore, the region exhibiting the lowest Pi value was identified in IR regions, indicating that the IR regions were highly conserved in Boraginaceae species.

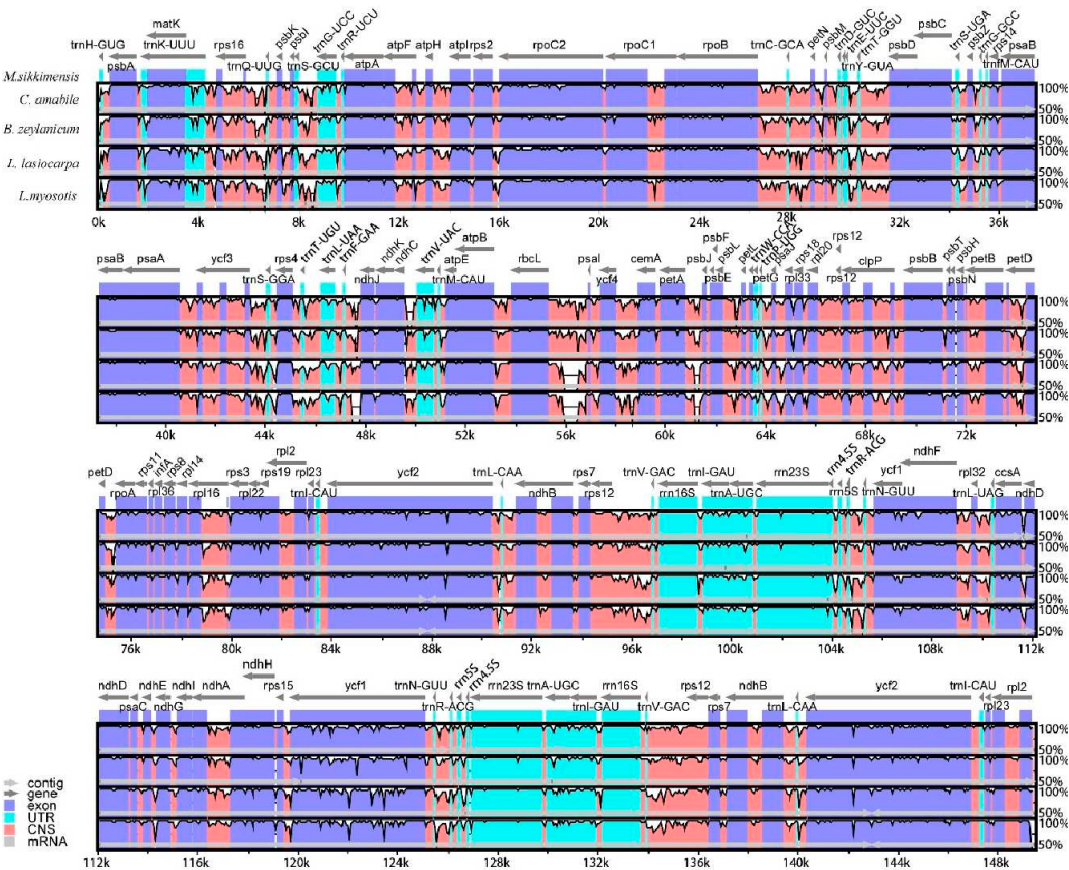


Figure 4. Sequence alignment of five Boraginaceae genomes in MVISTA. The grey arrows above the alignment indicate the genes transcription directions. The Y-axis represents identity, ranging from 50% to 100%.

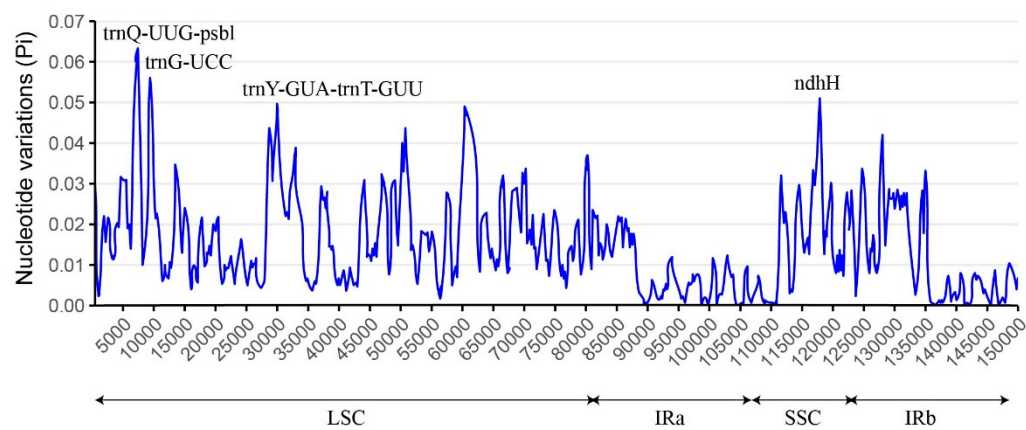


Figure 5. Nucleotide diversity (Pi) analysis for chloroplast genomes from the Boraginaceae plants. (window length: 600 bp; step size: 200 bp).

The expansion and contraction of the IR region is an important aspect of plant evolution, which can lead to structural changes in the chloroplast genome, affecting the expression and function of chloroplast genes [10,41,45]. We conducted a comparative analysis of the chloroplast genome region boundaries for the following five species: *M. sikkimensis*, *Bothriospermum zeylanicum*, *Trigonotis zhuokejiensis*, *Trigonotis tibetica*, and *Cynoglossum amabile*. The results showed that the length of the chloroplast genomes in the five species ranged from 148,193 bp (*T. tibetica*) to 152,532 bp (*C. amabile*). The LSC region displayed lengths from 80,767 bp (*T. tibetica*) to 83,692 bp (*B. zeylanicum*), while the length of SSC region spanned from 17,181 bp (*B. zeylanicum*) to 17,366 bp (*C. amabile*). Furthermore, the length of IR region ranged from 25,088 bp (*T. tibetica*) to 25,632 bp (*C. amabile*). Across all species, the boundary of LSC-IRb region and SSC-Ira region were found to be located in the *rps19* and *ycf1* genes, respectively. The junction between SSC and IRb was located in the *ycf1* gene in most species except *B. zeylanicum*, in which there was no *ycf1* gene near the IRb-SSC region but *ndhF*, which was completely encoded in SSC region and exhibited a 1bp distance to the junction of the IRb/SSC region. In *T. zhuokejiensis* and *T. tibetica* chloroplast genomes, the IRb-SSC boundary was situated within the *ndhF* gene, which had a 2 bp insertion in IRb. Furthermore, other genes of *rpl22*, *rpl2* and *trnH* were also found in the LSC/IR and SSC/IR boundary among the chloroplast genome from these five species (Figure 6).

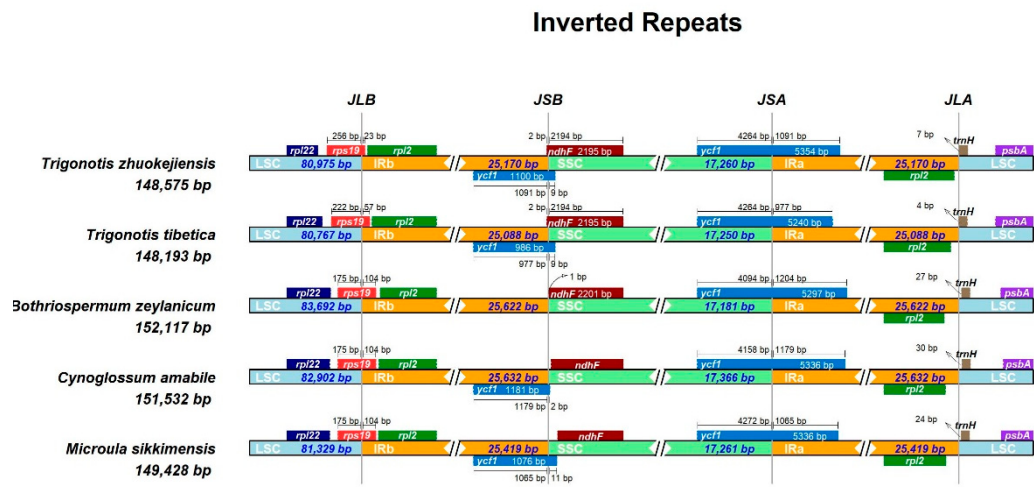


Figure 6. Comparison of the junction positions between the LSC, SSC and IR regions among the chloroplast genomes of five species.

3.5. Phylogenetic analysis

The phylogenetic trees were generated using ML and BI methods based on the chloroplast genomes of the 30 Boraginaceae species with the *Isodon serra* and *Forsythia suspensa* as the outgroups. As shown in Figure 7, most nodes showed high support values. These 30 Boraginaceae species could be classified into two subfamilies: Cynoglossoideae and Boraginoideae.

M.sikkimensis displayed a sister relationship with *C. amabile* and *B. zeylanicum*, but failed to cluster into the same branch, indicating obvious distinctions in their chloroplast genomes, albeit with high similarity. The present study provided indispensable genetic insights for clarifying the phylogenetic relationship of family Boraginaceae plants.

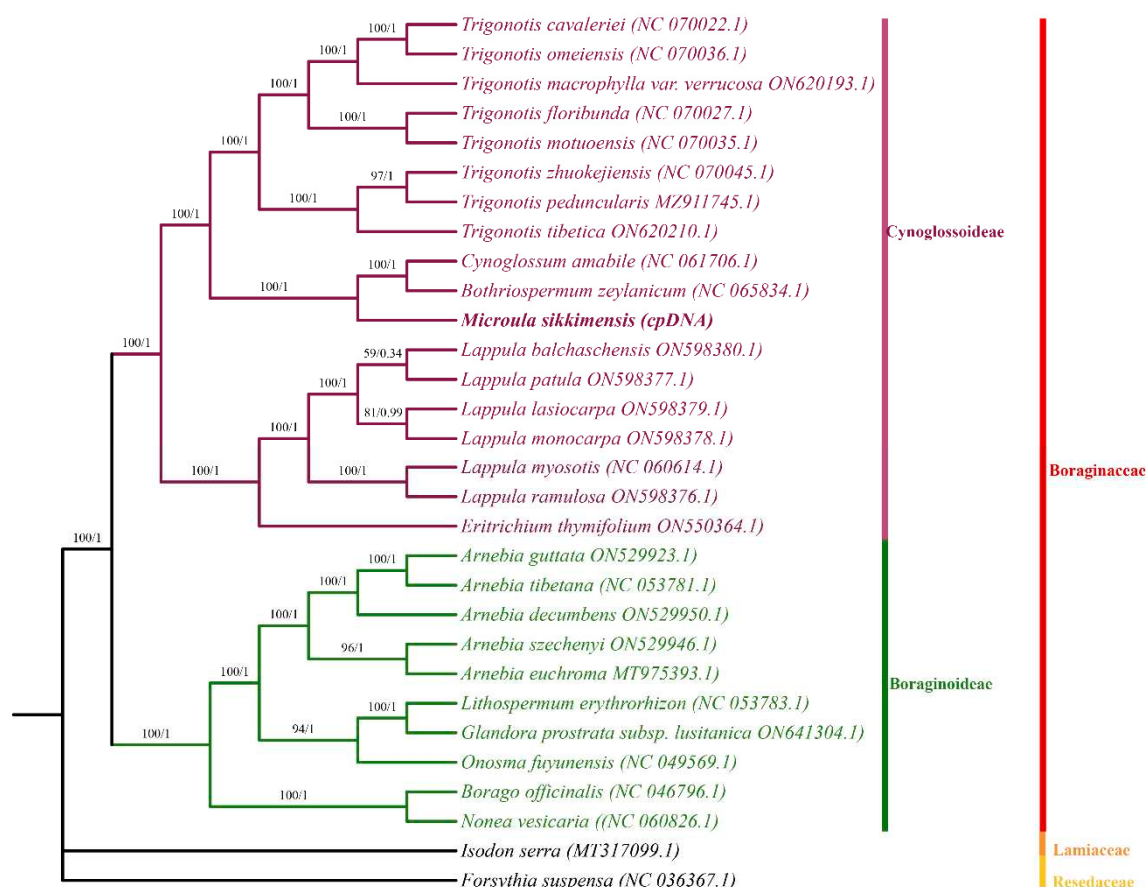


Figure 7. Phylogenetic tree reconstructed based on the complete chloroplast genome sequences from 30 species using ML method. The numerical annotations above the branches represented the ML bootstrap support values / BI probability support values.

4. Discussion

In this study, the complete chloroplast genome of *M. sikkimensis* was sequenced and its genetic information was reported for the first time. The results revealed that the chloroplast genome of *M. sikkimensis* spanned a length of 149,428 bp, exhibiting the characteristic circular quadripartite structure, comprising the LSC, SSC, and two IR regions. Gene annotation indicated that a total of 112 genes, including 78 protein-coding genes, 30 tRNA genes, and 4 rRNA genes were found in the *M. sikkimensis* chloroplast genome, which was consistent with the chloroplast genome structure and gene count in other Boraginaceae species [46–50]. Introns play an important role in the regulation of gene expression. They influence the synthesis and function of proteins through splicing regulation, thus significantly impacting the development, growth and environment adaptation of organisms [51].

Twelve intron-containing genes were identified in *M. sikkimensis*, including 10 genes with one intron and 2 genes with two introns.

SSRs, as a crucial category of molecular genetic markers, was found extensive applications across various domains of biological research. They served as important tools applied in genetic relationships, population structures, and evolutionary processes among species [52–54]. In plant populations, SSRs could serve as highly effective markers, facilitating the examination of genetic diversity within closely related taxa. Here, we found 38 SSRs in the chloroplast genome in *M. sikkimensis*, including five types of SSRs: mononucleotide, dinucleotide, trinucleotide, tetranucleotide, pentanucleotide. A/T mononucleotide repeat were the most frequent SSR, followed by AT dinucleotide repeat, that were consistent with the previous reports [55,56]. These SSR markers provided insights for genetic diversity studies and conservation strategies in the Boraginaceae.

Research on codon usage bias could explain gene expression and translation strategies. Synonymous codon usage bias was associated with intron number and was non-homogeneous across all exons. The pattern of its heterogeneity differed from species to species. It has also been shown that DNA methylation was likely a major driver of synonymous codon usage bias [57]. The GC content of codons considered as one of the primary factors contributing to codon usage bias formation [58]. Among the 30 codons with RSCU values higher than 1, 29 terminated with A or U. However, among the 32 codons with RSCU values lower than 1, 29 ended with G or C. This result indicated a preference in *M. sikkimensis* for codons with A or U as the third base, which was similar with *Fagopyrum dibotrys* and *Salix wilsonii* [59,60]. This preference for A or T might be a result of natural selection and gene mutation. In the chloroplast genome of plants, the majority of optimal synonymous codons ended with A or U, which might be due to the higher content of A and T, leading to the occurrence of A or T codon bias [61].

Through comparison with the chloroplast genomes of other species, deeper understanding of the evolutionary relationships and genomic structural disparities between these species could be gained [62,63]. Comparative analysis of the chloroplast genomes of five Boraginaceae species revealed substantial similarity in the coding regions, yet significant distinctions in the non-coding regions. These highly variable regions could serve as molecular markers for subsequent studies in Boraginaceae, holding vital significance for species classification and phylogenetic research. Additionally, phylogenetic analysis revealed that *M. sikkimensis* had a sister relationship with *C. amabile* and *B. zeylanicum*. However, they failed to cluster into the same branch, indicating that they were highly similar in chloroplast genomes, but there were still discernible differences.

5. Conclusion

This study presented a comprehensive analysis of the chloroplast genome of *M. sikkimensis*, shedding light on its genetic characteristics and evolutionary context within the Boraginaceae family. The chloroplast genome, spanning over 149,428 bp, exhibited the characteristic circular quadripartite structure common to higher plants. The gene annotation revealed a total of 112 genes, consistent with the structure observed in related Boraginaceae species. The identification of SSRs, with a predominance of A/T mononucleotide repeats followed by AT dinucleotide repeats, provided valuable genetic markers for exploring genetic diversity within closely related plant populations. Comparative analysis with other species chloroplast genomes revealed both conservation and variability, particularly in non-coding regions. These highly variable regions served as crucial molecular markers for species classification and phylogenetic research within the Boraginaceae family. Furthermore, phylogenetic analysis placed *M. sikkimensis* in a distinct relationship with *C. amabile* and *B. zeylanicum*, indicating a high degree of similarity in their chloroplast genomes. However, discernible differences existed, suggesting subtle genomic distinctions that endorse further investigation. Overall, this study not only provided a comprehensive understanding of the chloroplast genome of *M. sikkimensis* but also established a valuable genetic resource for future research in phylogenetics, molecular marker development, and conservation strategies within the Boraginaceae family. These findings could contribute to the broader field of plant genomics and facilitate a deeper appreciation of the evolutionary dynamics within this plant family.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org. Table S1 All information of species and the accession numbers of their chloroplast genomes in NCBI; Table S2 The intron-containing genes in the chloroplast genomes of *M. sikkimensis*; Table S3 Sequences of protein-coding genes of *M. sikkimensis* chloroplast genome; Table S4 repeat sequences in the chloroplast genome of *M. sikkimensis*.

Author Contributions: C.L. conceived and designed the experiments. Z.C., Y.G. and X.L. sampled plant specimens and conducted experiments. Y.G. performed data analyses. Y.G. wrote the manuscript. Z.C. and K.M. revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Major Science and Technology Project of Qinghai Province (2021-SF-A4); Chinese Academy of Sciences–People’s Government of Qinghai Province on Sanjiangyuan National Park (LHZX-2022-01); Gansu Province Grassland Monitoring and Evaluation Technology Support Project of Gansu Province Forestry and Grassland Administration((2021)794).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in NCBI (GenBank accession number: OR866440).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Han, F.; Cheng, D.Z.; Shi, S.B.; Ran, F.; Li, Y.K.; Bao, S.K.; Ren, F.; Shi, L.N.; Han, Q. The research process of a high - quality wild resource - *Microula sikkimensis* (in Chinese). *Chinese wild plant resources* **2007**, 5-9.
2. Zheng, S.Z.; Yang, H.P.; Meng, J.C.; Ma, X.M.; Shen, X.W. Studies on the constituents from the seeds of *M.sikkimensis* (in Chinese). *Journal of northwest normal university (natural science)* **2003**, 54-57, doi:10.16783/j.cnki.nwnuz.2003.02.017.
3. Wu, L.P. Exploitation and research of *Microula sikkimensis* resources (in Chinese). *China oils and fats* **1994**, 41-42.
4. Luo, G.R.; Deng, Y.C. Experiment on feeding *Microula sikkimensis* straw to Tibetan sheep. *Journal of grassland and forage science* **2001**, 56-57.
5. Szabò, I.; Spetea, C. Impact of the ion transportome of chloroplasts on the optimization of photosynthesis. *J Exp Bot* **2017**, 68, 3115-3128, doi:10.1093/jxb/erx063.
6. Mullineaux, P.M.; Exposito-Rodriguez, M.; Laissue, P.P.; Smirnoff, N. ROS-dependent signalling pathways in plants and algae exposed to high light: Comparisons with other eukaryotes. *Free radical biology & medicine* **2018**, 122, 52-64, doi:10.1016/j.freeradbiomed.2018.01.033.
7. Pollari, M.; Ruotsalainen, V.; Rantamäki, S.; Tyystjärvi, E.; Tyystjärvi, T. Simultaneous inactivation of sigma factors B and D interferes with light acclimation of the cyanobacterium *Synechocystis* sp. strain PCC 6803. *Journal of bacteriology* **2009**, 191, 3992-4001, doi:10.1128/jb.00132-09.
8. Yang, X.; Xie, D.F.; Chen, J.P.; Zhou, S.D.; Yu, Y.; He, X.J. Comparative Analysis of the Complete Chloroplast Genomes in *Allium Subgenus Cyathophora* (Amaryllidaceae): Phylogenetic Relationship and Adaptive Evolution. *BioMed research international* **2020**, 2020, 1732586, doi:10.1155/2020/1732586.
9. Yang, J.B.; Tang, M.; Li, H.T.; Zhang, Z.R.; Li, D.Z. Complete chloroplast genome of the genus *Cymbidium*: lights into the species identification, phylogenetic implications and population genetic analyses. *BMC evolutionary biology* **2013**, 13, 84, doi:10.1186/1471-2148-13-84.
10. Daniell, H.; Lin, C.S.; Yu, M.; Chang, W.J. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome biology* **2016**, 17, 134, doi:10.1186/s13059-016-1004-2.
11. Kim, G.B.; Lim, C.E.; Kim, J.S.; Kim, K.; Lee, J.H.; Yu, H.J.; Mun, J.H. Comparative chloroplast genome analysis of *Artemisia* (Asteraceae) in East Asia: insights into evolutionary divergence and phylogenomic implications. *BMC genomics* **2020**, 21, 415, doi:10.1186/s12864-020-06812-7.
12. Olmstead, R.G.; Palmer, J.D. Chloroplast DNA systematics: a review of methods and data analysis. *American Journal of Botany* **1994**, 81, 1205-1224.
13. Wang, A.; Wu, H.; Zhu, X.; Lin, J. Species Identification of *Conyza bonariensis* Assisted by Chloroplast Genome Sequencing. *Front Genet* **2018**, 9, 374, doi:10.3389/fgene.2018.00374.

14. Kelchner, S. The Evolution of Non-Coding Chloroplast DNA and Its Application in Plant Systematics. *Annals of the Missouri Botanical Garden* **2000**, *87*, 482-498, doi:10.2307/2666142.
15. Tang, J.; Xia, H.; Cao, M.; Zhang, X.; Zeng, W.; Hu, S.; Tong, W.; Wang, J.; Wang, J.; Yu, J.; et al. A comparison of rice chloroplast genomes. *Plant Physiol* **2004**, *135*, 412-420, doi:10.1104/pp.103.031245.
16. Li, J.; Tang, J.; Zeng, S.; Han, F.; Yuan, J.; Yu, J. Comparative plastid genomics of four *Pilea* (Urticaceae) species: insight into interspecific plastid genome diversity in *Pilea*. *BMC Plant Biol* **2021**, *21*, 25, doi:10.1186/s12870-020-02793-7.
17. Saski, C.; Lee, S.B.; Daniell, H.; Wood, T.C.; Tomkins, J.; Kim, H.G.; Jansen, R.K. Complete chloroplast genome sequence of *Gycine max* and comparative analyses with other legume genomes. *Plant Mol Biol* **2005**, *59*, 309-322, doi:10.1007/s11103-005-8882-0.
18. Du, Q.; Li, J.; Wang, L.; Chen, H.; Jiang, M.; Chen, Z.; Jiang, C.; Gao, H.; Wang, B.; Liu, C. Complete chloroplast genomes of two medicinal *Swertia* species: the comparative evolutionary analysis of *Swertia* genus in the Gentianaceae family. *Planta* **2022**, *256*, 73, doi:10.1007/s00425-022-03987-z.
19. Daniell, H.; Lee, S.-B.; Grevich, J.; Saski, C.; Quesada-Vargas, T.; Guda, C.; Tomkins, J.; Jansen, R.K. Complete chloroplast genome sequences of *Solanum bulbocastanum*, *Solanum lycopersicum* and comparative analyses with other Solanaceae genomes. *Theoretical and Applied Genetics* **2006**, *112*, 1503-1518, doi:10.1007/s00122-006-0254-x.
20. Shi, L.; Chen, H.; Jiang, M.; Wang, L.; Wu, X.; Huang, L.; Liu, C. CPGAVAS2, an integrated plastome sequence annotator and analyzer. *Nucleic acids research* **2019**, *47*, W65-w73, doi:10.1093/nar/gkz345.
21. Greiner, S.; Lehwark, P.; Bock, R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic acids research* **2019**, *47*, W59-w64, doi:10.1093/nar/gkz238.
22. Lowe, T.M.; Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research* **1997**, *25*, 955-964, doi:10.1093/nar/25.5.955.
23. Chen, Y.; Ye, W.; Zhang, Y.; Xu, Y. High speed BLASTN: an accelerated MegaBLAST search tool. *Nucleic acids research* **2015**, *43*, 7762-7768, doi:10.1093/nar/gkv784.
24. Liu, S.; Ni, Y.; Li, J.; Zhang, X.; Yang, H.; Chen, H.; Liu, C. CPGView: A package for visualizing detailed chloroplast genome structures. *Molecular ecology resources* **2023**, *23*, 694-704, doi:10.1111/1755-0998.13729.
25. Lewis, S.E.; Searle, S.M.; Harris, N.; Gibson, M.; Lyer, V.; Richter, J.; Wiel, C.; Bayraktaroglu, L.; Birney, E.; Crosby, M.A.; et al. Apollo: a sequence annotation editor. *Genome biology* **2002**, *3*, Research0082, doi:10.1186/gb-2002-3-12-research0082.
26. Frazer, K.A.; Pachter, L.; Poliakov, A.; Rubin, E.M.; Dubchak, I. VISTA: computational tools for comparative genomics. *Nucleic acids research* **2004**, *32*, W273-279, doi:10.1093/nar/gkh458.
27. Amiryousefi, A.; Hyvönen, J.; Poczai, P. IRscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics (Oxford, England)* **2018**, *34*, 3030-3031, doi:10.1093/bioinformatics/bty220.
28. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* **2013**, *30*, 772-780, doi:10.1093/molbev/mst010.
29. Rozas, J.; Ferrer-Mata, A.; Sánchez-DelBarrio, J.C.; Guirao-Rico, S.; Librado, P.; Ramos-Onsins, S.E.; Sánchez-Gracia, A. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Molecular biology and evolution* **2017**, *34*, 3299-3302, doi:10.1093/molbev/msx248.
30. Zhang, D.; Gao, F.; Jakovlić, I.; Zou, H.; Zhang, J.; Li, W.X.; Wang, G.T. PhyloSuite: An integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. *Molecular ecology resources* **2020**, *20*, 348-355, doi:10.1111/1755-0998.13096.
31. Kumar, S.; Stecher, G.; Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular biology and evolution* **2016**, *33*, 1870-1874, doi:10.1093/molbev/msw054.
32. Beier, S.; Thiel, T.; Münch, T.; Scholz, U.; Mascher, M. MISA-web: a web server for microsatellite prediction. *Bioinformatics (Oxford, England)* **2017**, *33*, 2583-2585, doi:10.1093/bioinformatics/btx198.
33. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* **1999**, *27*, 573-580, doi:10.1093/nar/27.2.573.
34. Kurtz, S.; Choudhuri, J.V.; Ohlebusch, E.; Schleiermacher, C.; Stoye, J.; Giegerich, R. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic acids research* **2001**, *29*, 4633-4642, doi:10.1093/nar/29.22.4633.

35. Talavera, G.; Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic biology* 2007, 56, 564-577, doi:10.1080/10635150701472164.
36. Nguyen, L.T.; Schmidt, H.A.; von Haeseler, A.; Minh, B.Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution* 2015, 32, 268-274, doi:10.1093/molbev/msu300.
37. Ronquist, F.; Teslenko, M.; van der Mark, P.; Ayres, D.L.; Darling, A.; Höhna, S.; Larget, B.; Liu, L.; Suchard, M.A.; Huelsenbeck, J.P. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology* 2012, 61, 539-542, doi:10.1093/sysbio/sys029.
38. Letunic, I.; Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic acids research* 2019, 47, W256-w259, doi:10.1093/nar/gkz239.
39. Yi, D.K.; Kim, K.J. Two complete chloroplast genome sequences of genus *Paulownia* (Paulowniaceae): *Paulownia coreana* and *P. tomentosa*. *Mitochondrial DNA. Part B, Resources* 2016, 1, 627-629, doi:10.1080/23802359.2016.1214546.
40. Zheng, L.P.; Li, L.J. Characterization of the complete chloroplast genome of *Centranthera grandiflora* Benth (Orobanchaceae), an important species of medicinal herb. *Mitochondrial DNA. Part B, Resources* 2021, 6, 1784-1785, doi:10.1080/23802359.2021.1934139.
41. Shen, X.; Wu, M.; Liao, B.; Liu, Z.; Bai, R.; Xiao, S.; Li, X.; Zhang, B.; Xu, J.; Chen, S. Complete Chloroplast Genome Sequence and Phylogenetic Analysis of the Medicinal Plant *Artemisia annua*. *Molecules (Basel, Switzerland)* 2017, 22, doi:10.3390/molecules22081330.
42. Zhang, F.; Chen, H.; Zhou, Y.; Li, N.; Chong, X.; Li, Y.; Lu, X.; Wang, C. The complete chloroplast genome sequence and phylogenetic analysis of *Ilex 'Beryl'*, a hybrid of *Ilex cornuta* × *Ilex latifolia* (Aquifoliaceae). *Mitochondrial DNA. Part B, Resources* 2021, 6, 227-228, doi:10.1080/23802359.2020.1861569.
43. Zeng, S.; Zhou, T.; Han, K.; Yang, Y.; Zhao, J.; Liu, Z.L. The Complete Chloroplast Genome Sequences of Six Rehmannia Species. *Genes (Basel)* 2017, 8, doi:10.3390/genes8030103.
44. Nie, L.; Cui, Y.; Wu, L.; Zhou, J.; Xu, Z.; Li, Y.; Li, X.; Wang, Y.; Yao, H. Gene Losses and Variations in Chloroplast Genome of Parasitic Plant *Macrosolen* and Phylogenetic Relationships within Santalales. *Int J Mol Sci* 2019, 20, doi:10.3390/ijms20225812.
45. Xue, S.; Shi, T.; Luo, W.; Ni, X.; Iqbal, S.; Ni, Z.; Huang, X.; Yao, D.; Shen, Z.; Gao, Z. Comparative analysis of the complete chloroplast genome among *Prunus mume*, *P. armeniaca*, and *P. salicina*. *Horticulture research* 2019, 6, 89, doi:10.1038/s41438-019-0171-1.
46. Sun, J.; Wang, S.; Wang, Y.; Wang, R.; Liu, K.; Li, E.; Qiao, P.; Shi, L.; Dong, W.; Huang, L.; et al. Phylogenomics and Genetic Diversity of *Arnebia Radix* and Its Allies (Arnebia, Boraginaceae) in China. *Front Plant Sci* 2022, 13, 920826, doi:10.3389/fpls.2022.920826.
47. Zhao, F.; Peng, H. The complete chloroplast genome of *Caryopteris incana* (Lamiaceae) and phylogenetic analysis. *Mitochondrial DNA Part B* 2020, 5, 1399-1400, doi:10.1080/23802359.2020.1736963.
48. Duan, H.C.; Zheng, X.H.; Li, Y.Y.; Li, S.M.; Ye, L.; Jing, H.Z.; Dong, Q. The complete chloroplast genome of *Fraxinus malacophylla* (Oleaceae, Oleoideae). *Mitochondrial DNA. Part B, Resources* 2020, 5, 3588-3589, doi:10.1080/23802359.2020.1827999.
49. He, Y.; Xu, X.; Liu, Q. The complete chloroplast genome of *Onosma fuyunensis* Y. He & Q.R. Liu and its phylogenetic analysis. *Mitochondrial DNA. Part B, Resources* 2021, 6, 3142-3143, doi:10.1080/23802359.2020.1861567.
50. Wu, J.H.; Li, H.M.; Lei, J.M.; Liang, Z.R. The complete chloroplast genome sequence of *Trigonotis peduncularis* (Boraginaceae). *Mitochondrial DNA. Part B, Resources* 2022, 7, 456-457, doi:10.1080/23802359.2022.2048212.
51. Wu, L.; Fan, P.; Zhou, J.; Li, Y.; Xu, Z.; Lin, Y.; Wang, Y.; Song, J.; Yao, H. Gene Losses and Homology of the Chloroplast Genomes of *Taxillus* and *Phacellaria* Species. *Genes (Basel)* 2023, 14, doi:10.3390/genes14040943.
52. Ebert, D.; Peakall, R. Chloroplast simple sequence repeats (cpSSRs): technical resources and recommendations for expanding cpSSR discovery and applications to a wide array of plant species. *Molecular ecology resources* 2009, 9, 673-690, doi:10.1111/j.1755-0998.2008.02319.x.
53. George, B.; Bhatt, B.S.; Awasthi, M.; George, B.; Singh, A.K. Comparative analysis of microsatellites in chloroplast genomes of lower and higher plants. *Current genetics* 2015, 61, 665-677, doi:10.1007/s00294-015-0495-9.

54. Du, X.; Zeng, T.; Feng, Q.; Hu, L.; Luo, X.; Weng, Q.; He, J.; Zhu, B. The complete chloroplast genome sequence of yellow mustard (*Sinapis alba* L.) and its phylogenetic relationship to other Brassicaceae species. *Gene* 2020, 731, 144340, doi:10.1016/j.gene.2020.144340.
55. Mwanzia, V.M.; Nzei, J.M.; Yan, D.Y.; Kamau, P.W.; Chen, J.M.; Li, Z.Z. The complete chloroplast genomes of two species in threatened monocot genus *Caldesia* in China. *Genetica* 2019, 147, 381-390, doi:10.1007/s10709-019-00079-x.
56. Dong, S.; Ying, Z.; Yu, S.; Wang, Q.; Liao, G.; Ge, Y.; Cheng, R. Complete chloroplast genome of *Stephania tetrandra* (Menispermaceae) from Zhejiang Province: insights into molecular structures, comparative genome analysis, mutational hotspots and phylogenetic relationships. *BMC genomics* 2021, 22, 880, doi:10.1186/s12864-021-08193-x.
57. Qin, Z.; Cai, Z.; Xia, G.; Wang, M. Synonymous codon usage bias is correlative to intron number and shows disequilibrium among exons in plants. *BMC genomics* 2013, 14, 56, doi:10.1186/1471-2164-14-56.
58. Zhang, Z.; Dai, W.; Wang, Y.; Lu, C.; Fan, H. Analysis of synonymous codon usage patterns in torque teno sus virus 1 (TTSuV1). *Archives of Virology* 2013, 158, 145-154, doi:10.1007/s00705-012-1480-y.
59. Wang, X.; Zhou, T.; Bai, G.; Zhao, Y. Complete chloroplast genome sequence of *Fagopyrum dibotrys*: genome features, comparative analysis and phylogenetic relationships. *Sci Rep* 2018, 8, 12379, doi:10.1038/s41598-018-30398-6.
60. Chen, Y.; Hu, N.; Wu, H. Analyzing and Characterizing the Chloroplast Genome of *Salix wilsonii*. *BioMed research international* 2019, 2019, 5190425, doi:10.1155/2019/5190425.
61. Zhang, H.; Huang, T.; Zhou, Q.; Sheng, Q.; Zhu, Z. Complete Chloroplast Genomes and Phylogenetic Relationships of *Bougainvillea spectabilis* and *Bougainvillea glabra* (Nyctaginaceae). *International Journal of Molecular Sciences* 2023, 24, 13044.
62. Jiao, L.; Lu, Y.; He, T.; Li, J.; Yin, Y. A strategy for developing high-resolution DNA barcodes for species discrimination of wood specimens using the complete chloroplast genome of three *Pterocarpus* species. *Planta* 2019, 250, 95-104, doi:10.1007/s00425-019-03150-1.
63. Bi, Y.; Zhang, M.F.; Xue, J.; Dong, R.; Du, Y.P.; Zhang, X.H. Chloroplast genomic resources for phylogeny and DNA barcoding: a case study on *Fritillaria*. *Sci Rep* 2018, 8, 1184, doi:10.1038/s41598-018-19591-9.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.