Analysis of Six Chloroplast Genomes Provides Insight into the Evolution of Chrysosplenium (Saxifragaceae)

Zhihua Wu
South-Central University for Nationalities
Rui Liao
South-Central University for Nationalities
Tiange Yang
South-Central University for Nationalities
Xiang Dong
Wuhan Botanical Garden
Deqing Lan
South-Central University for Nationalities
Rui Qin
South-Central University for Nationalities
Hong Liu (✉ liuhong@mail.scuec.edu.cn)
South-central University for Nationalities https://orcid.org/0000-0001-7227-2476

Research article

Keywords: Saxifragaceae, Chrysosplenium, Chloroplast genome, Opposite leaves, Alternate leaves, Phylogenomics

Posted Date: June 2nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-26814/v1

Version of Record: A version of this preprint was published on September 10th, 2020. See the published version at https://doi.org/10.1186/s12864-020-07045-4.

Abstract

Background

Chrysosplenium L. (Saxifragaceae) is a genus of plants widely distributed in China and usually found in moist, shaded valleys and mountain slopes. This genus is ideal for studying plant adaptation to low-light conditions. Although some progress has been made in the systematics and biogeography of *Chrysosplenium*, its chloroplast genome evolution remains to be investigated.

Results

To fill this gap, we sequenced the chloroplast genomes of six *Chrysosplenium* species and analyzed their genome structure, GC content, and nucleotide diversity. Moreover, we performed a phylogenetic analysis and calculated non-synonymous (Ka)/synonymous (Ks) substitution ratios using the combined protein-coding genes of 29 species within Saxifragales and two additional species as outgroups, as well as a pair-wise estimation for each gene within *Chrysosplenium*. Compared with the outgroups in Saxifragaceae, the six *Chrysosplenium* chloroplast genomes had lower GC contents; they also had conserved boundary regions and gene contents, as only the *rpl32* gene was lost in four of the *Chrysosplenium* chloroplast genomes. Phylogenetic analyses suggested that the *Chrysosplenium* separated to two major clades (the opposite group and the alternate group). The pair-wise Ka/Ks ratios of genes in the *Chrysosplenium* species showed that *matK* and *ycf2* were subjected to relatively relaxed selection.

Conclusion

This study provides genetic resources for exploring the phylogeny of *Chrysosplenium* and sheds light on plant adaptation to low-light conditions. The lower average GC content and the lacking gene of *rpl32* indicating selective pressure in their unique habitats. Different from results previously reported, our selective pressure estimation suggested that the genes related to photosynthesis (such as *ycf2*) were under positive selection at sites in the coding region.

1 Background

Chrysosplenium L. is a genus of Saxifragaceae and belongs to the subfamily Saxifragoideae according to the APG IV [1]. The genus is important in the taxonomy of Saxifragaceae due to its unique characters: *Chrysosplenium* flowers have four sepals, lack petals, and have two equal or distinctly unequal fruiting capsules [2]. The genus plays an important role in the phylogeny of Saxifragaceae, comprises about 79 perennial herbs [3], and mainly occurs in the northern hemisphere, with the highest species diversity in East Asia; only two species, *Chrysosplenium valdivicum* Hook and *Chrysosplenium macranthum* Hook, are found in the Southern Hemisphere [4–7]. Thirty-six species and fifteen variants have been recorded in China [6–9]. *Chrysosplenium* is divided into two subgenera according to the leaf arrangement: *Alternifolia* Franchet with alternate leaves and *Oppositifolia* Franchet with opposite leaves [4, 8, 10]. Some species within the genus have been used in traditional Tibetan medicine to treat digestive system diseases [11]. The genus was regarded as a typical group with floristic disjunction and is important for studying speciation [12, 13].

Nuclear markers (such as the internal transcribed spacer, ITS) of the ribosomal DNA, and chloroplast markers were employed to determine the molecular phylogeny of *Chrysosplenium* [14, 15]. Compared to nuclear markers, the chloroplast genome possesses highly conserved DNA sequences and a lower substitution level (especially in inverted repeat regions). Therefore, the chloroplast genome is ideal for phylogenetic inference at the species and higher levels [5, 15–18]. The first *Chrysosplenium* chloroplast genome has been sequenced and assembled [19]; additional complete chloroplast genomes are required for investigating the evolutionary features of *Chrysosplenium*.

Challenging environments may impose selective pressure on genes, which could leave a footprint of natural selection in genes involved in adaptation to the environment. For example, low light conditions have a major effect on the evolution of the chloroplast genome. Marcelino et al. (2016) analyzed the chloroplast genome of *Ostreobium quekettii*, a green alga that lives in extremely low light environments. Surprisingly, the genes related to photosynthesis were found to be under strong purifying selection, rather than the expected positive selection. The study also found that the low light niche may contribute to the reduction of genome size. This study contributed to our understanding of plant adaptive evolution; however, to our knowledge, a comparative analysis of chloroplast genomes of angiosperms with low-light requirements has not been conducted. *Chrysosplenium* (Saxifragaceae) is usually found in shaded valleys and mountain slopes and, compared with other genera within Saxifragaceae, this genus has the lowest light requirement [4, 20]. Therefore, examination of the chloroplast genomes of *Chrysosplenium* species may provide insight into the effect of low light in angiosperms.

In this study, we aimed to provide a comprehensive insight into the evolution of the chloroplast genomes of several *Chrysosplenium* species. First, we sequenced the chloroplast genomes of six *Chrysosplenium* species. Next, we conducted comparative chloroplast genome analyses for these six genomes, plus four Saxifragaceae chloroplast genomes from GenBank. Then, we constructed a phylogeny of *Chrysosplenium* using chloroplast genomes of 29 species within Saxifragales and two outgroups. Finally, we estimated selective pressures to investigate whether the genes related to photosynthesis in *Chrysosplenium* are under purifying selection or positive selection.

2 Results

2.1 Organization of the Chloroplast Genomes of *Chrysosplenium* Species

The chloroplast genomes of the *Chrysosplenium* species contain the typical quadripartite structures (Fig. 1), which include long single copy (LSC), inverted repeat (IR) and small single copy (SSC) regions. The seven *Chrysosplenium* chloroplast genomes ranged from 151,679 bp to 153,460 bp in length (see Table 1 for details), with IRs 25,974–26,224 bps, LSCs 82,771–83,752 bps and SSCs 16,960–17,342 bps. Each *Chrysosplenium* chloroplast genomes encoded 30 transfer RNAs (tRNAs) and 4 ribosomal RNAs (rRNAs). Each genome also includes 79 functional proteins encoding genes except for *C. macrophyllum*, *C. flagelliferum*, *C. alternifolium*, *C. ramosum*, which lacked *rpl32*. In total, each chloroplast genome includes 113 (*rpl32* present) or 112 (*rpl32* loss) genes. The *rps12* gene in *Chrysosplenium* was recognized as a *trans*-spliced gene, with the first exon located in the LSC region and the other one or two exons distributed in the IR regions. In addition, 17 intron-containing genes were also detected (Supplementary Table S4). The chloroplast genome size and gene content neither significantly diverged between subg. *Oppositifolia* and *Alternifolia* (Table 1) nor significantly diverged between *Chrysosplenium* and other genera of Saxifragaceae.

Table 1
General information and comparison of chloroplast genomes of Saxifragaceae species

Characteristic	***C. macrophyllum***	***C. flagelliferum***	***C. alternifolium***	***C. kamtschaticum***	***C. ramosum***	***C. sinicum***	***C. aureobracteatum***
Size (base pair, bp)	152,837	151,679	152,619	152,561	153,460	153,427	153,102
LSC length (bp)	83,583	82,771	83,524	83,175	83,670	83,745	83,753
SSC length (bp)	17,264	16,960	17,111	16,986	17,342	17,236	17,317
IR length (bp)	25,995	25,974	25,992	26,200	26,224	26,223	26,016
Number of genes	112	112	112	113	112	113	113
Protein-coding genes	78	78	78	79	78	79	79
rRNA genes	4	4	4	4	4	4	4
tRNA genes	30	30	30	30	30	30	30
LSC GC%	35.33	35.26	35.35	35.28	35.24	35.05	35.20
SSC GC%	31.42	31.37	31.40	31.46	31.64	31.27	31.16
IR GC%	42.89	42.87	42.86	42.71	42.69	42.75	42.85
Lacking gene	*rpl32*	*rpl32*	*rpl32*		*rpl32*		

2.2 GC Content, Nucleotide Diversity, and Repeat Analysis

When we compared the total GC content of the chloroplast genomes of *Chrysosplenium* species with that of the chloroplast genomes of the three non-Chrysosplenium Saxifragaceae species (*S. stolonifera*, *B. scopulosa*, and *O. rupifraga*), we found the *Chrysosplenium* species have the lowest total GC contents (<37.5%) (Figure 2 and Supplementary Table S5). In addition, *Chrysosplenium* has the lowest GC contents (<29.7%) at the third codon position (GC3). Within the *Chrysosplenium* species, the GC contents in subgenera *Oppositifolia* were slightly lower than those in subgenera *Alternifolia*, regardless of the total GC contents or those in GC3.

The IR regions were more conserved than the LSC and SSC regions, with average Pi values of 0.00586 in IR regions, 0.01760 in the LSC region, and 0.019 in the SSC region (Supplementary Figure S1 and Supplementary Table S6). In the LSC region, *psbT* has the highest Pi value of 0.22159, followed by *trnG*–GCC with Pi value of 0.10369.

Among the mono-, di-, tri-, tetra-, penta-, and hexa-nucleotide categories of SSRs in the chloroplast genomes of the *Chrysosplenium* species, mono-nucleotide repeats were the most common (Supplementary Table S7 and Supplementary Figure S2A) ranging from 42.42% (*C. sinicum*) to 61.29% (*C. flagelliferum*). Hexa-nucleotide repeats account for the lowest proportion of SSRs in *C. ramosum*, *C. sinicum*, and *C. flagelliferum*. *Chrysosplenium* species contained fewer SSRs than *B. scopulosa* and *O. rupifraga*. Among the four repeat types, the most common repeat type was palindromic repeats, which ranged from 53.13% in *C. aureobracteatum* to 42.42% in *C. macrophyllum* (Supplementary Table S8 and Supplementary Figure S2 B).

2.3 Boundary Regions and Comparative Analysis

When comparing the chloroplast genomes of *Chrysosplenium* species, we found that IR–LSC junctions of IRb are largely located between *rpl2* and *rps19* (Figure 3). Moreover, the overlap of *ycf1* pseudogenes and *ndhF* appeared in different locations in the *Chrysosplenium* species: in the region of the SSC for *C. ramosum*, and at the IRb–SSC border for the other five species. *C.*

alternifolium did not contain the *ycf1* pseudogene. The *ycf1* gene was sited at the SSC–IRa boundary and the length of *ycf1* ranged from 5,402–5,546 bps. The *trnH* gene of the seven *Chrysosplenium* species was located in the LSC region, 2–19 bp away from the IRa–LSC border.

When comparing the genome boundaries of the *Chrysosplenium* species to the other three non-Chrysosplenium species of Saxifragaceae, *ndhF* was at the IRb–SSC boundary in most species of *Chrysosplenium*, except for *C. ramosum*, which showed contraction of the SSC and expansion of IRb. In addition, *S. stolonifera* was slightly different from the other two non-Chrysosplenium species in Saxifragaceae. In *S. stolonifera*, the contraction of the LSC region resulted in the *rpl22* gene being at the IRb–LSC junction, which placed the whole *rps19* gene in the IRb region. The *rps19* pseudogenes were also found in the IRa region in *S. stolonifera* and *B. scopulosa*. When these data were combined with the phylogenetic tree of the three clades (non-*Chrysosplenium*, *Alternifolia*, and *Oppositifolia*) inferred from whole-chloroplast protein-coding genes (Figure 3), we found that the chloroplast genome structure within *Chrysosplenium* species is not strongly conserved, although the gene content is conserved.

LAGAN and Shuffle-LAGAN gave very similar results in the genetic divergence among the chloroplast genomes of Saxifragaceae species (Figure 4). The chloroplast genomes of the *Chrysosplenium* species were more conserved when compared with the three non-Chrysosplenium species of Saxifragaceae, and the intergenic spacer (IGS) regions had the highest levels of divergence: *trnK*–*rps16*, *rps16*–*trnQ*, *rpoB*–*trnC*, *petN*–*psbM*, *trnT*–*psbD*, *psbZ*–*trnG*, *trnT*–*trnL*, *accD*–*psaI*, *ycf4*–*cemA*, *ndhF*–*rpl32*, and *rps15*–*ycf1*. In addition, we found some highly variable coding sequences (*ndhD*, *ycf2*, *ndhA*, and *ycf1*), and the IR regions were more conserved than LSC and SSC regions in all the species tested. We also found slight difference for *rpoC2*, *ycf2*, and *ycf1*, which correspond to the difference between the *Alternifolia* and *Oppositifolia* subgenera.

2.4 Selective pressure analyses

We calculated the Ka/Ks ratios at the species level by concatenating all of the 79 genes into a super-matrix. In *Chrysosplenium* species, the Ka/Ks ratios were around 0.2. This result suggested that at the whole-chloroplast protein level, the *Chrysosplenium* species have been subjected to a stronger purifying selection (Figure 5, Supplementary Table S9 and Supplementary Table S10).

The Ka/Ks ratio was also calculated for all of the 79 protein-coding genes of the ten chloroplast genomes of *Chrysosplenium* separately (Figure 6 and Supplementary Table S7). Two genes (*matK*, *ycf2*) had Ka/Ks ratios around 1.0 in most species, implying relatively relaxed selection. Specially, *matK* showed an average Ka/Ks ratio of 0.74 when compared with *C. ramosum*. Among the *Chrysosplenium* species, *ycf2* often had a ratio higher than 0.8. Most of the other genes had a Ka/Ks ratio range from 0.1–0.3, implying strongly purification (Table S10).

Sixty-six single-copy genes were used for selective pressure estimation with the branch-site model (Supplementary Table S9). We found that *matK* was positively selected in *Chrysosplenium* with the p-value = 0.022 and the BEB probability for one amino acid site larger than 0.972. In addition, positively selected sites were detected for 18 genes (*atpB*, *atpE*, *atpF*, *atpI*, *cemA*, *clpP*, *matK*, *ndhC*, *ndhE*, *ndhF*, *ndhH*, *ndhK*, *petA*, *psaB*, *psbH*, *psbJ*, *psbN*, *rps14*, *rps16*) (Figure 7). Among them, each gene had mutations that led to changes in the amino acid sequence (see Supplementary Table S9 for details).

2.5 Phylogenetic Analysis

Phylogenetic analyses yielded a well-supported phylogeny of Saxifrageles with most of the nodes having ML bootstrap support values >95 and BI posterior probabilities =1 (Figure 8). The topologies yielded from ML analysis and BI analysis were completely identical. The topology of Saxifragales in our study was similar to the APG □ system [1] with Saxifragaceae grouped with Iteaceae. *Chrysosplenium* was divided into two clades corresponding to the two subgenera (*Alternifolia* and *Oppositifolia*).

3 Discussion

3.1 *rpl32* gene was lost in four of the seven *Chrysosplenium* species

The six chloroplast genomes of *Chrysosplenium* shared a typical quadripartite structure with similar genome sizes and gene composition. However, gene numbers were slightly different due to the loss of *rpl32* or transfer of the gene to the nucleus [21, 22]. *rpl32* was not present in the chloroplast of four species, viz.: *C. macrophyllum*, *C. flagelliferum*, *C. alternifolium*, and *C. ramosu*m. In *Populus alba* and *Thalictrum coreanum*, the missing chloroplast *rpl32* gene has been identified in the nuclear genome [23-25]. The gene transfer could be explained by the decreased demand on photosynthesis and plastid translational capacity, which increased the success rate of gene transfer to the nucleus [26]. The *Chrysosplenium* species always grow in moist, shaded valleys and mountain slopes, which may influence photosynthesis. Moreover, the *rpl32* gene is usually lost in the *Alternifolia*. In a word, the loss or transfer of *rpl32* may be a *Chrysosplenium* adaptation to their special habitats, although this hypothesis remains to be verified by experiments.

3.2 Lower GC content was detected in the chloroplast genomes of *Chrysosplenium*

Compared to other non-Chrysosplenium members of the Saxifragaceae, the chloroplast genomes of *Chrysosplenium* have an overall low GC content, which can be explained by the natural selection or neutral mutation. According to Foerstner et al. (2005), DNA sequences of closely related species from different environments show marked differences in GC content, which has a direct impact on the amino acid sequences of the proteins in the respective environments. Genes with low GC contents are more prone to be transcribed than those with high GC content as GC pairs have three hydrogen bonds, making them more stable than AT pairs with two hydrogen bonds. Therefore, the selective pressure of the unique habitat of *Chrysosplenium* species (insufficient light energy) resulted in the lower overall GC contents and GC3 contents in their chloroplast genomes.

3.3 Strong purifying selection was detected for most of the chloroplast genes in *Chrysosplenium* species

Chrysosplenium is usually found in moist, shaded valleys and mountain slopes. Compared with other genera within Saxifragaceae, the genus has the lowest light requirement. Among the 79 chloroplast genes in plants, 46 are related to photosynthesis pathway [27]. Genes related to a specific environment are normally assumed to be under positive selection [28]. This assumption was widely used to detect genes related to environmental adaptation [29]. Our expectation was that the 46 genes were under positive selection.

However, the lower Ka/Ks ratios at the chloroplast genome level within the *Chrysosplenium* species compared to non-*Chrysosplenium* species indicated that most genes were subjected to purifying selection to retain conserved functions in the *Chrysosplenium*. In the opposite environment, sunlight, including UV radiation, induces DNA damage, mutations and rearrangements [29, 30], which may contribute to an increase in mutation rates. Moreover, it has been proposed that more solar radiation and higher temperatures increase metabolism and growth rates, shortening generation times and increasing mutation rates [31]. We could speculate that low-light lineages will likely have slower rates of molecular evolution than lineages living in high-light conditions.

3.4 *matK* was under positive selection

The gene *matK* is transcribed from the sole intact plastid group IIA intron ORF localized between the exons coding for the lysine-tRNA (*trnK-UUU*). In contrast to other group IIA ORFs, *matK* has lost domains assigned to a reverse transcriptase and endonuclease function [27, 32]. *matK* is usually used as a phylogenetic signal that can resolve evolutionary relationships because of its the high nucleotide and amino acid substitution rates [33, 34].

The molecular evolution of the *matK* coding region is unusual [34], and the results of a previous study differed from our result that showed an extremely significant positive selection site at the 117S loci (0.972*, from polar Ser to non-polar Val). Plants have a variety of strategies to adapt to the environment; therefore, multiple genes may have been subjected to positive selection in *Chrysosplenium* during its adaptation. Further functional studies on the adaptive amino acid sites of chloroplast genes in *Chrysosplenium* are needed.

4 Conclusions

In this study, we sequenced the chloroplast genomes of six *Chrysosplenium* species and revealed the chloroplast genomic features between the *Oppositifolia* (*C. macrophyllum*, *C. flagelliferum*, *C. alternifolium*, and *C. ramosum*) and *Alternifolia* (*C. ramosum*, *C. kamtschaticum*, and *C. sinicum*) subgenera. In addition, we combined these six sequences with the previously reported chloroplast genomes of *C. aureobracteatum* (*Oppositifolia*), *S. stolonifera* (Saxifragaceae), *B. scopulosa* (Saxifragaceae), and *O. rupifraga* (Saxifragaceae). We discussed the comprehensive features of the chloroplast genomes, such as gene content and GC content, in these seven species of *Chrysosplenium*. All the species of Saxifragaceae shared similar genome structures, whereas the seven species of *Chrysosplenium* showed a lower average GC content, indicating selective pressure in their unique habitats. At the chloroplast genome level, the Ka/Ks ratios of the individual sequences showed that *Chrysosplenium* species were subjected to purifying selection compared to the non-*Chrysosplenium* species. We found that *matK* and *ycf2* were possibly under weak purifying selection. For the *Chrysosplenium* genus, many genes involved in photosynthesis were subjected to positive selection via a change of amino acid sites, which may be the adaptive response to its moist and shaded habitat. For example, 15 of 19 genes involved in photosynthesis showed positive selection with significant posterior probabilities. Using the protein-coding sequences from the whole chloroplast genome of 31 species, the robust consensus of phylogenetic trees reconstructed with both ML and BI algorithms suggested that *Chrysosplenium* species are sister to *B. scopulosa* and *O. rupifraga* within Saxifragaceae of Saxifragales. Also, our results supported the classification of the genus into two subgenera based on the morphology of opposite leaves or alternate leaves. These findings will be valuable for further study of the chloroplast genomes of *Chrysosplenium* species and provide valuable resources for studies of plant adaptation to low-light conditions.

6 Materials And Methods

6.1 Sampling and sequencing

To represent the *Chrysosplenium*, six species were selected based on their morphological characteristics: *C. macrophyllum*, *C. flagelliferum*, and *C. alternifolium* belonging to the *Alternifolia* subgenera, and *C. ramosum*, *C. kamtschaticum*, and *C. sinicum* belonging to the *Oppositifolia* subgenera. Among the six species from wild, three were collected from China and three were from Japan (See details in Supplementary Table S1). All these six species were formal identified by Deqing Lan. Due to a high content of secondary metabolites, the chloroplast DNA of *C. macrophyllum* was extracted using a high-salt method [35]. To get a complete chloroplast genome, which can be used as a reference in assembling chloroplast genome of the other five species, *C. macrophyllum* was sequenced using the PacBio Sequel I platform at Frasergen (Wuhan, China) and the Illumina Hiseq 2500 at the Novogene Company (Beijing, China). The total genomic DNA for the other five *Chrysosplenium* species was extracted using a modified cetyltrimethylammonium bromide (CTAB) method [36] and sequenced using the Illumina Hiseq 2500 platform at the Novogene Company (Beijing, China).

6.2 Chloroplast Genome Assembly and Annotation

The sequencing of the chloroplast DNA of *C. macrophyllum* with the PacBio Sequel I platform generated 218,330 reads with the N50 of 4,452 bp. *De novo* genome assembly was conducted using Canu (v1.5) [37], which produced 4,028 contigs with an N50 of 5,011 bp. To discard nuclear DNA sequences, we aligned the contigs to a whole-chloroplast reference genome with the Burrow-Wheeler Aligner bwa [38]. Then the contigs were polished with Arrow implemented in SMRT Link v6.0.0. Finally, the draft chloroplast genome was manually adjusted based on the two inverted repeats and scaffolds assembled from the Illumina Hiseq 2500 platform.

Sequence data for the other five species generated from the Illumina Hiseq 2500 platform were processed to remove the low-quality reads and adaptors. The clean reads were aligned to the complete chloroplast genome of *C. macrophyllum* with bwa-0.7.12 [38]. The aligned reads were then selected for *de novo* assembly with ABYSS-2.0.2 [39] after the optimal Kmer was chosen with the software kmergenie [40]. Then, the contigs were connected with Sequencher 5.4.6 and scaffolded again with the original data by the software SSPACE_Standard_v3.0 [41]. Last, the assembled scaffolds were manually adjusted based on the two inverted repeats and verified by Sanger sequencing (Supplementary Table S2). We also assembled these scaffolds with GetOrganelle [42] to validate the ones assembled with ABYSS-2.0.2.

Gene annotation was performed using CPGAVAS2 [43] and PGA [44]. The different annotations of protein-coding sequences were confirmed using BLASTx. tRNAs were checked with tRNAscan-SE v2.0.3 [45]. Final chloroplast genome maps were drawn using OGDRAW [46].

6.3 Analysis of GC Content, Nucleotide Diversity, and Repeat Content

We accessed chloroplast genome sequences of *Chrysosplenium aureobracteatum* (NC_039740; Saxifragaceae; *Chrysosplenium*), *Saxifraga stolonifera* (NC_037882; Saxifragaceae; *Saxifraga*), *Bergenia scopulosa* (NC_036061; Saxifragaceae; *Bergenia*), and *Oresitrophe rupifraga* (NC_037514; Saxifragaceae; *Oresitrophe*) from GenBank to compare the features among these chloroplast genomes in Saxifragaceae. The nucleotide diversity (Pi) among the seven species of *Chrysosplenium* was calculated using the software DnaSP v6.12.03 [47]. The GC content of the whole chloroplast genome and the third position GC content of codons for all ten species were calculated using an in-house Python script. The simple sequence repeats (SSRs) were detected using the MIcroSAtellite (MISA) identification tool with the minimum repeat number set at 10, 5, 4, 3, 3, and 3 for mono-, di-, tri-, tetra-, penta-, and hexanucleotides, respectively. We also identified tandem repeat sequences using REPuter [48] with minimal repeats of more than 30 bp and hamming distances of less than 3 bp.

6.4 Boundary Regions and Comparative Analysis

The contraction or expansion between boundary regions of the chloroplast genome in each species was drawn by IRscope [49]. To compare the conservation of each gene, we visualized the results with mVISTA through two alignment programs: LAGAN, which produces true multiple alignments regardless of whether they contain inversions or not, and Shuffle-LAGAN, which can detect rearrangements and inversions in sequences [50].

6.5 Selective pressure estimation

We carried out selective pressure estimation for the 6 species of *Chrysosplenium* and 25 species of non-Chrysosplenium with two strategies: calculation based on pairwise comparison and calculation based on the branch-site model (Yang et al., 2015):

The *Chrysosplenium* genome includes 79 genes. The protein-coding sequences (CDSs) of the 79 genes from each of the 31 species were concatenated into a super matrix. Then, species vs. species Ka/Ks ratio was estimated. In addition, the Ka/Ks ratio was estimated for each of the 79 genes within *Chrysosplenium* separately. The CDS for each gene was translated to amino acid sequences, which were aligned with MEGA7 (Kumar et al., 2016). Then, the corresponding CDS were aligned according to the amino acid sequences. Lastly, Ka/Ks ratios were calculated using the KaKs-calculator v 2.0 (Wang et al., 2010). Genes with 1 < Ka/Ks ratio < 45 were considered as under positive selection; genes with Ka/Ks ratio < 1 were considered as under purifying selection. The ratio >=45 or NA indicates that the gene has few nonsynonymous sites/substitutions, and was not considered in our analysis.

A preliminary analysis found that 66 single-copy chloroplast genes were not lost or transferred to the nucleus. The 66 genes were used for analyses using the branch-site model [28]. CDSs of each gene were aligned according to their amino acid sequences with MEGA7 (Kumar et al., 2016). The branch-site model in the program codeml within the PAML v4.9 package (Yang, 2007) was used to assess potential positive selection in *Chrysosplenium* that was set as the foreground branch. Selective pressure is measured by the ratio (ω) of the nonsynonymous substitution rate (dN) to the synonymous substitutions rate (dS). A neutral branch-site model (Model = 2, NSsites = 2, Fix = 1, and Fix ω = 1) and an alternative branch-site model (Model = 2, NSsites= 2, and Fix = 0) were applied separately. The right-tailed Chi-square test was used to compute p-values based on the difference of log-likelihood values between the two models with one degree of freedom. Moreover, Bayes Empirical Bayes (BEB) method (Yang et al., 2015) was implemented to calculate the posterior probabilities for amino acid sites that are potentially under positive selection. A gene with a p-value < 0.05 and ω >1 was considered as a positively selected gene. An amino acid site with posterior probabilities >0.95 was considered as positively selected.

6.6 Phylogenetic Analyses

To construct a phylogeny of *Chrysosplenium*, 29 species of Saxifrageles and two Buxaceae species (as outgroups) were selected (see Supplementary Table S3 for details). The whole-chloroplast protein-coding genes of these 31 species were aligned with MUSCLE v3.8.31 [51]. The best-fitting nucleotide substitution model was determined using the Akaike Information Criterion in the model-finder IQ-TREE [52]. Maximum likelihood analysis was performed using IQ-TREE with the best model of GTR+F+R4 and 1000 bootstrap replicates, and Bayesian inference analysis was performed in MrBayes 3.2.6 [53] using the Markov Chain Monte Carlo method with 200,000 generations and sampling trees every 100 generations. The first 20% of trees were discarded as burn-in with the remaining trees being used for generating a consensus tree.

7. Declarations

7.1 Ethics approval and consent to participate

Not applicable.

7.2 Consent for publication

Not applicable.

7.3 Availability of data and material

All sequences used in this study can be found in Genbank according to their Genbank accession. The other data sets generated in this study are included within the article and supplementary files. All materials used or generated during the study are kept in our laboratory.

7.4 Competing interests

Not applicable. All authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

7.5 Funding

This work was supported by the Construction Plan of Hubei Province Science and Technology basic conditions platform (No.2017BEC014) and Fund for Key Laboratory Construction of Hubei Province (No.2018BFC360). All funders we mentioned provide financial support for our study.

7.6 Authors' Contributions

ZW, RL, and HL conceived and designed the experiments. DL contributed to the sampling. RL, TY and XD performed the experiments. ZW and LR analyzed the data. ZW, RL, and HL wrote the paper. All authors have read and approved the manuscript.

7.7 Acknowledgements

Not applicable.

8. References

1. The Angiosperm Phylogeny G, Chase MW, Christenhusz MJM, Fay MF, Byng JW, Judd WS, Soltis DE, Mabberley DJ, Sennikov AN, Soltis PS, et al. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Bot J Linn Soc. 2016;181(1):1–20.

2. Maximowicz CJ: **Diagnoses plantarum novarum Japoniae et Mandshuriae**, vol. 17. St. Petersburg, sér. 3: BuLletin de l'Academie Imperiale des Sciences de Saint-Pétersbourg; 1872.
3. POWO. **Plants of the World Online**. In. Facilitated by the Royal Botanic Gardens, Kew: Published on the Internet; 2019: http://powo.science.kew.org/taxon/urn:lsid:ipni.org:names:30129301-30129302#children.
4. Hara H. **Synopsis of the genus Chrysosplenium L**, vol. 7; 1957.
5. Soltis DE, Tago-Nakazawa M, Xiang Q-Y, Kawano S, Murata J, Wakabayashi M, Hibsch-Jetter C. Phylogenetic relationships and evolution in Chrysosplenium (Saxifragaceae) based on matK sequence data. Am J Bot. 2001;88(5):883–93.
6. Lan D, Qin R, Xia J, Li G, Liu H. New records of Chrysosplenium spp. in Wuling mountain area of the west of Hubei and Hunan. Biotic Resource. 2018;40(2):159–63.
7. Lan D, Liu H, Huang W, Yi L, Zhang D, Qin E, Qin R. Investigation and Analysis of Chrysosplenium Resources in Tibetan-inhabited Area. Journal of Plant Genetic Resources. 2019;20(3):662–8.
8. Pan JT, Ohba H. Chrysosplenium L. Vol. 8. Beijing: Science Press & Missouri Botanical Garden Press; 2001.
9. Liu H, Luo J-L, Liu Q-Y, Lan D-Q, Qin R, Yu X-LJP: **A new species of Chrysosplenium** (**Saxifragaceae**) **from Zhangjiajie**, **Hunan**, **central China**. 2016, **277**(3):287–292.
10. Franchet AR. **Monographie du genere Chrysosplenium Tourn**, vol. 3(2): Nouvelles Archives duMuséum d'Histoire Naturelle; 1890–1891.
11. Luo Y, Yu H, Yang Y, Tian W, Dong K, Shan J, Ma X. A flavonoid compound from Chrysosplenium nudicaule inhibits growth and induces apoptosis of the human stomach cancer cell line SGC-7901. Pharm Biol. 2016;54(7):1133–9.
12. Guo Q, Ricklefs RE, Cody ML. Vascular plant diversity in eastern Asia and North America: historical and ecological explanations. Bot J Linn Soc. 2008;128(2):123–36.
13. Xiang Q-Y, Zhang WH, Ricklefs RE, Qian H, Chen ZD, Wen J, Li JH. Regional Differences in Rates of Plant Speciation and Molecularevolution: A Comparison Between Eastern Asia and Eastern North America. Evolution. 2004;58(10):2175–84.
14. Han J-W, Yang S-G, Kim H-J, Jang C, Park J-M, Kang S-H. **Phylogenetic Study of Korean** *Chrysosplenium* **Based on nrDNA ITS Sequences**. 2011, 24(**4**):358–369.
15. Kim Y-I, Kim Y-D. Molecular Systematic Study of Chrysosplenium Series Pilosa (Saxifragaceae) in Korea. Journal of Plant Biology. 2011;54(6):396.
16. Ravi V, Khurana JP, Tyagi AK, Khurana PJPS. Evolution: **An update on chloroplast genomes**. 2008, 271(1/2):101–122.
17. Parks M, Cronn R, Liston A. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biol. 2009;7(1):84.
18. Xie D-F, Yu H-X, Price M, Xie C, Deng Y-Q, Chen J-P, Yu Y, Zhou S-D, He X-J: **Phylogeny of Chinese Allium Species in Section Daghestanica and Adaptive Evolution of Allium** (**Amaryllidaceae**, **Allioideae**) **Species Revealed by the Chloroplast Complete Genome**. 2019, **10**:460.
19. Kim Y-I, Lee J-H, Kim Y-D. The complete chloroplast genome of a Korean endemic plant Chrysosplenium aureobracteatum Y.I. Kim & Y.D. Kim (Saxifragaceae). Mitochondrial DNA Part B. 2018;3(1):380–1.
20. Ibrahim RK, De Luca V, Khouri H, Latchinian L, Brisson L, Charest PM. Enzymology and compartmentation of polymethylated flavonol glucosides in chrysosplenium americanum. Phytochemistry. 1987;26(5):1237–45.
21. Huang CY, Ayliffe MA, Timmis JN. Direct measurement of the transfer rate of chloroplast DNA into the nucleus. Nature. 2003;422(6927):72–6.
22. Stegemann S, Hartmann S, Ruf S, Bock R: **High-frequency gene transfer from the chloroplast genome to the nucleus**. *Proceedings of the National Academy of Sciences* 2003, **100**(15):8828.
23. Ueda M, Fujimoto M, Arimura S-i, Murata J, Tsutsumi N, Kadowaki K-i: **Loss of the rpl32 gene from the chloroplast genome and subsequent acquisition of a preexisting transit peptide within the nuclear gene in Populus**. *Gene* 2007, **402**(1):51–56.
24. Park S, Jansen RK, Park S. Complete plastome sequence of Thalictrum coreanum (Ranunculaceae) and transfer of the rpl32 gene to the nucleus in the ancestor of the subfamily Thalictroideae. BMC Plant Biol. 2015;15(1):40.
25. Cusack BP, Wolfe KH. When gene marriages don't work out: divorce by subfunctionalization. Trends Genet. 2007;23(6):270–2.

26. Fleischmann TT, Scharff LB, Alkatib S, Hasdorf S, Schöttler MA, Bock R. Nonessential Plastid-Encoded Ribosomal Proteins in Tobacco: A Developmental Role for Plastid Translation and Implications for Reductive Genome Evolution. Plant Cell. 2011;23(9):3137.

27. Wicke S, Schneeweiss GM, dePamphilis CW, Müller KF, Quandt D. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol. 2011;76(3):273–97.

28. Yang Z, Wong WSW, Nielsen R. Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection. Mol Biol Evol. 2005;22(4):1107–18.

29. Raven JA, Beardall J, Larkum AW, Sánchez-Baracaldo P. Interactions of photosynthesis with genome size and function. Philosophical Transactions of the Royal Society B: Biological Sciences. 2013;368(1622):20120264.

30. Kumar RA, Oldenburg DJ, Bendich AJ. Changes in DNA damage, molecular integrity, and copy number for plastid DNA and mitochondrial DNA during maize development. J Exp Bot. 2014;65(22):6425–39.

31. Rohde K. **Latitudinal gradients in species diversity: the search for the primary cause**. *Oikos* 1992:514–527.

32. Liere K, Link G. RNA-binding activity of the matK protein encodecd by the chloroplast trnk intron from mustard (Sinapis alba L.). Nucleic Acids Res. 1995;23(6):917–21.

33. Hilu KW, Borsch T, Müller K, Soltis DE, Soltis PS, Savolainen V, Chase MW, Powell MP, Alice LA. Evans RJAjob: **Angiosperm phylogeny based on matK sequence information**. 2003, 90(12):1758–1776.

34. Hilu KW. Liang g: **The matK gene: sequence variation and application in plant systematics**. Am J Bot. 1997;84(6):830–9.

35. Shi C, Hu N, Huang H, Gao J, Zhao Y-J, Gao L-Z. An improved chloroplast DNA extraction procedure for whole plastid genome sequencing. PloS one. 2012;7(2):e31468–8.

36. Doyle J: **DNA Protocols for Plants**. In: *Molecular Techniques in Taxonomy.* Edited by Hewitt GM, Johnston AWB, Young JPW. Berlin, Heidelberg: Springer Berlin Heidelberg; 1991: 283–293.

37. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome research. 2017;27(5):722–36.

38. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589–95.

39. Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G, Khan H, Coombe L, Warren RL, et al. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. Genome research. 2017;27(5):768–77.

40. Medvedev P, Chikhi R. Informed and automated k-mer size selection for genome assembly. Bioinformatics. 2013;30(1):31–7.

41. Henkel CV, Butler D, Jansen HJ, Boetzer M, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 2010;27(4):578–9.

42. Jin J-J, Yu W-B, Yang J-B, Song Y, Yi T-S, Li D-Z. **GetOrganelle: a simple and fast pipeline for de novo assembly of a complete circular chloroplast genome using genome skimming data**. *bioRxiv* 2018:256479.

43. Shi L, Chen H, Jiang M, Wang L, Wu X, Huang L, Liu C. CPGAVAS2, an integrated plastome sequence annotator and analyzer. Nucleic Acids Res. 2019;47(W1):W65–73.

44. Qu X-J, Moore MJ, Li D-Z, Yi T-S. PGA: a software package for rapid, accurate, and flexible batch annotation of plastomes. Plant Methods. 2019;15(1):50.

45. Schattner P, Brooks AN, Lowe TM. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. 2005;33(suppl_2):W686–9.

46. Greiner S, Lehwark P, Bock R. **OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes**. *bioRxiv* 2019:545509.

47. Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sánchez-Gracia A. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. Mol Biol Evol. 2017;34(12):3299–302.

48. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 2001;29(22):4633–42.

49. Amiryousefi A, Hyvönen J, Poczai P. IRscope: an online program to visualize the junction sites of chloroplast genomes. Bioinformatics. 2018;34(17):3030–1.

50. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S. research NCSPJG: **LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA**. 2003, 13(4):721–731.

51. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.

52. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular biology evolution. 2015;32(1):268–74.

53. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. Syst Biol. 2012;61(3):539–42.

Figures

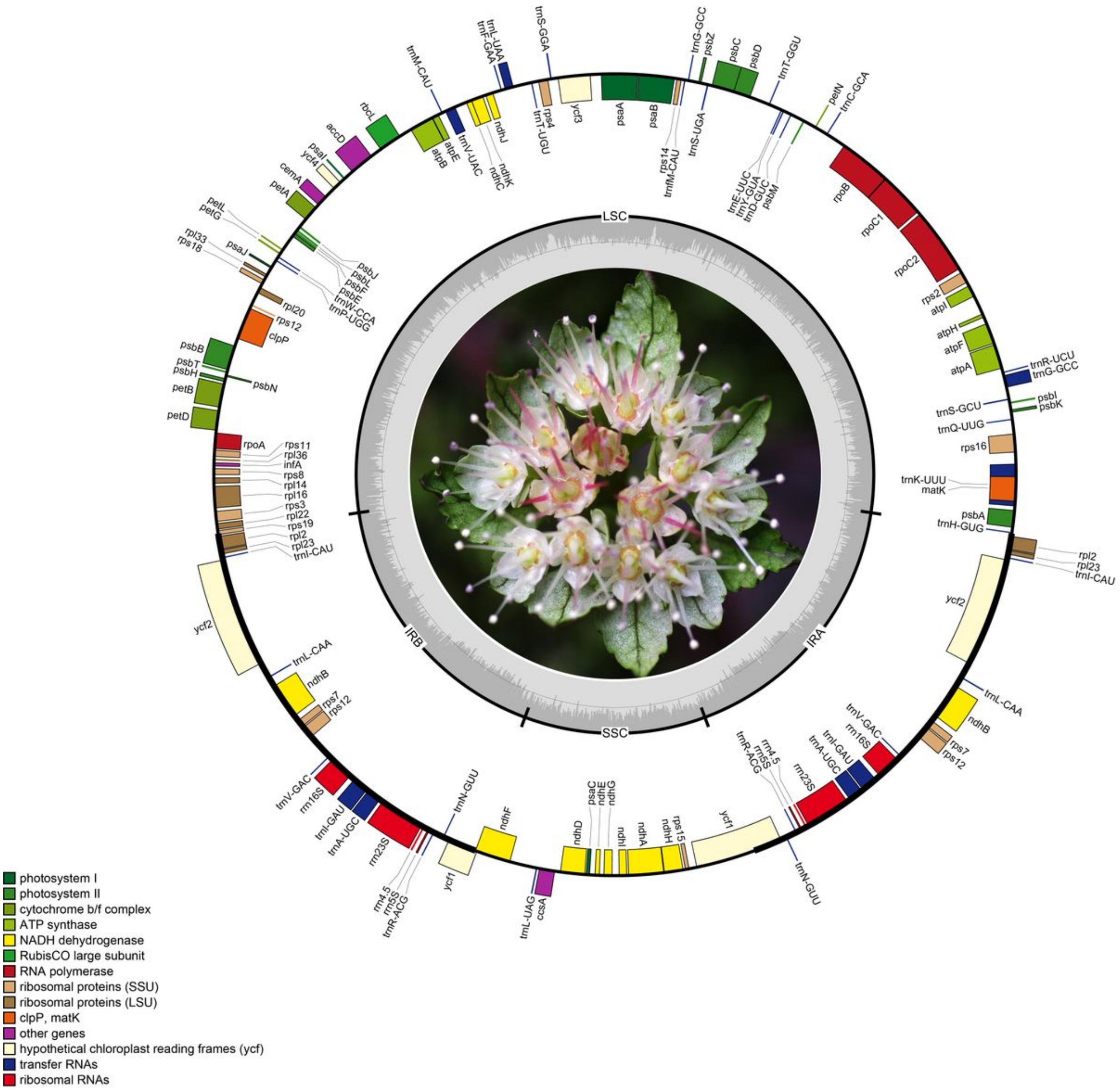

Figure 1

Gene map of the Chrysosplenium macrophyllum chloroplast genomes. Genes inside the circle are transcribed clockwise, genes outside are transcribed counter-clockwise. Genes are color-coded to indicate functional groups. The dark gray area in the inner circle corresponds to GC content while the light gray corresponds to the adenine-thymine (AT) content of the genome. The small (SSC) and large (LSC) single copy regions and inverted repeat (IRa and IRb) regions are noted in the inner circle.

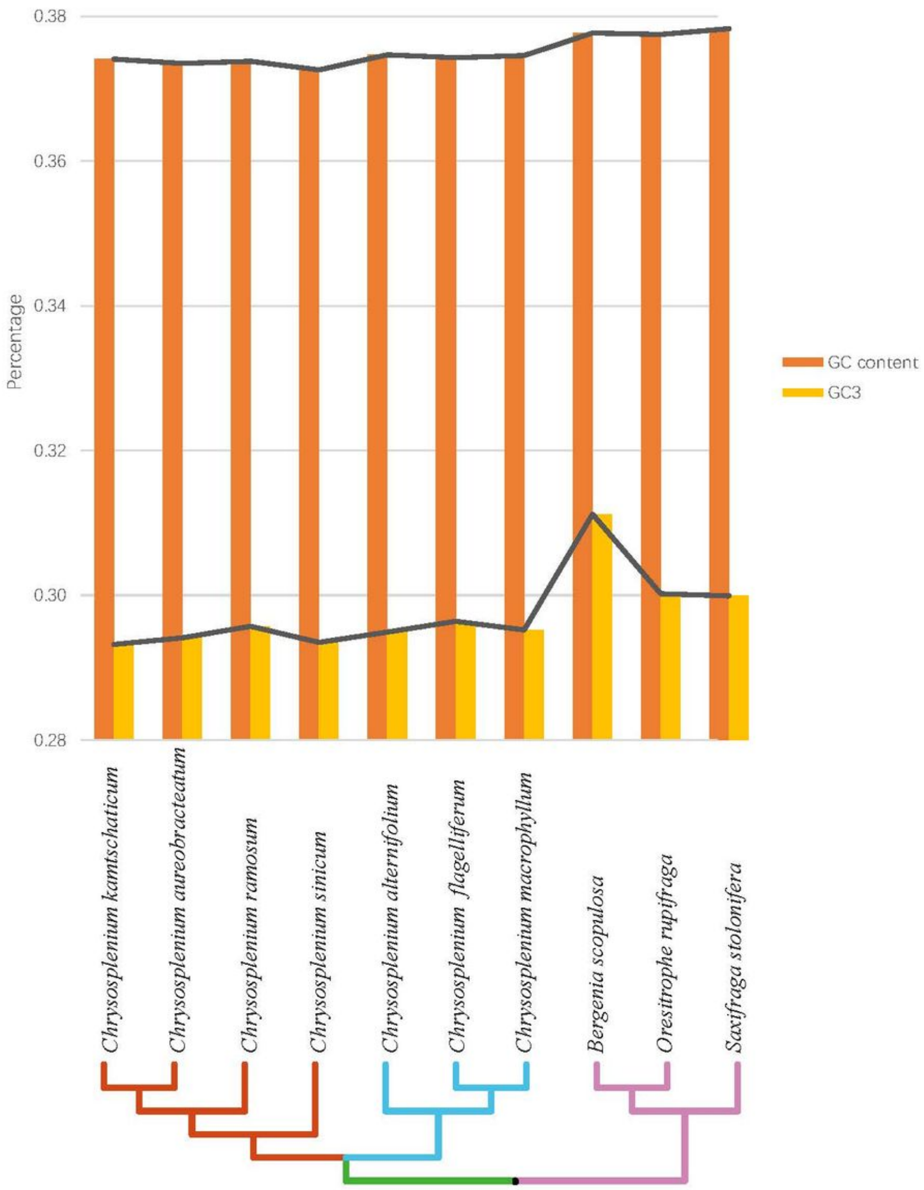

Figure 2

Changes in plastid GC content of Saxifragaceae. This graph shows the total GC content (orange bar and black line) and the third codon position GC content (yellow bar and gray line) of each species.

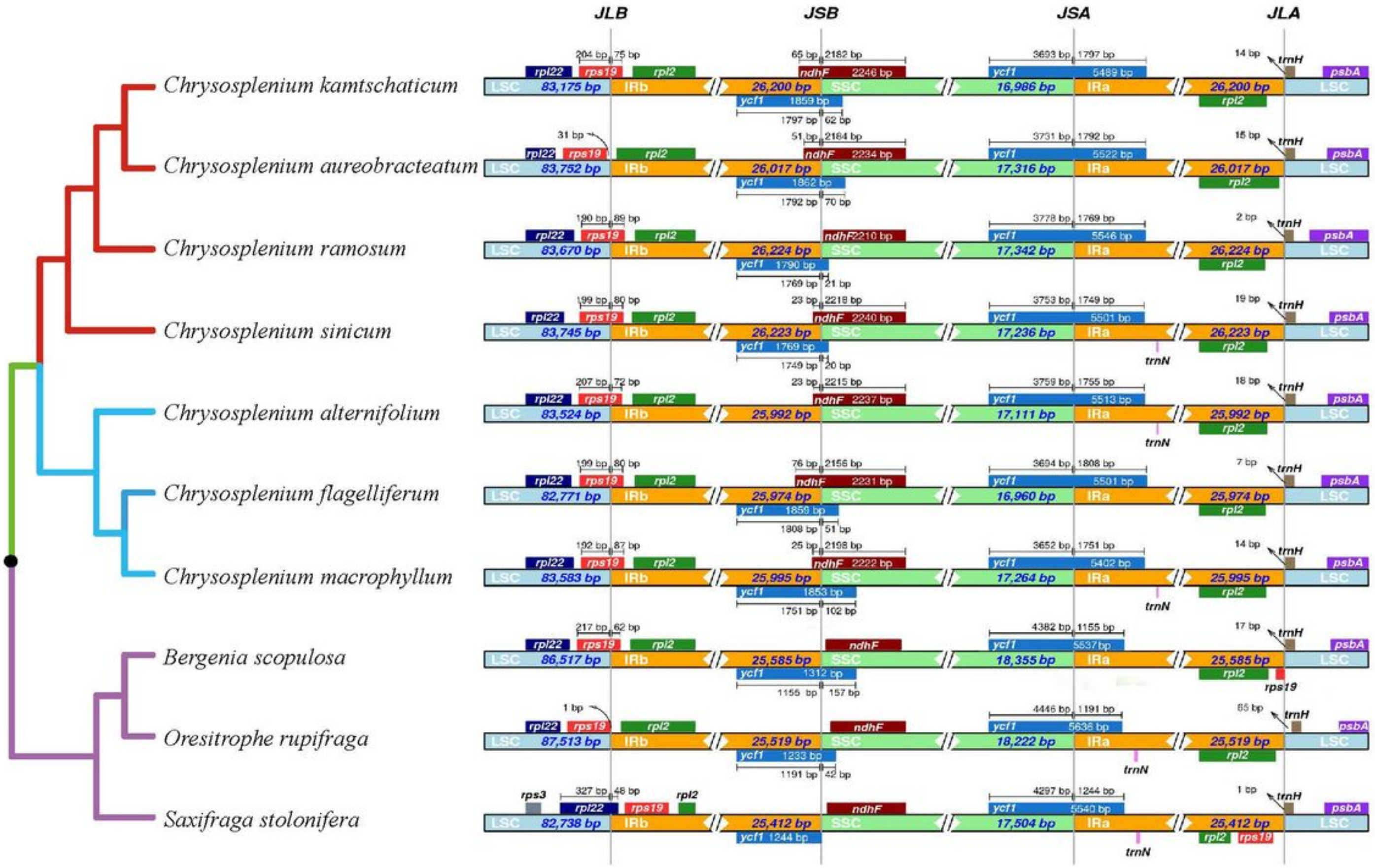

Figure 3

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

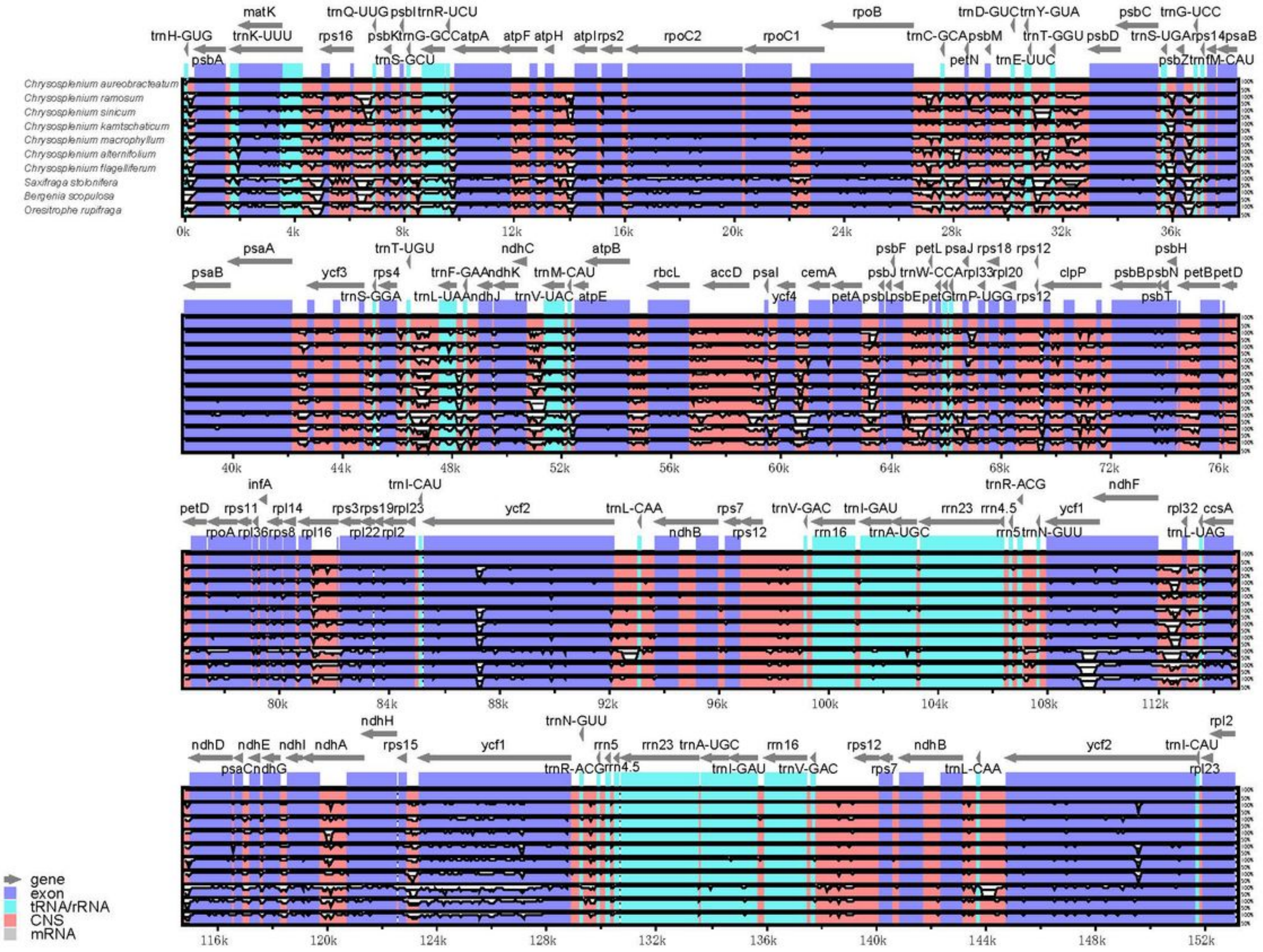

Figure 4

The comparative analysis with LAGAN program of the whole-chloroplast genome of seven different species from the family of Saxifragaceae. The percentage of identity is shown in the vertical axis, ranging from 50% to 100%, while the horizontal axis shows the position within the chloroplast genome. Each arrow displays the annotated genes and direction of their transcription in the reference genome (C. aureobracteatum). Genome regions are color-coded as exon, tRNA, conserved noncoding sequences (CNS), and mRNA.

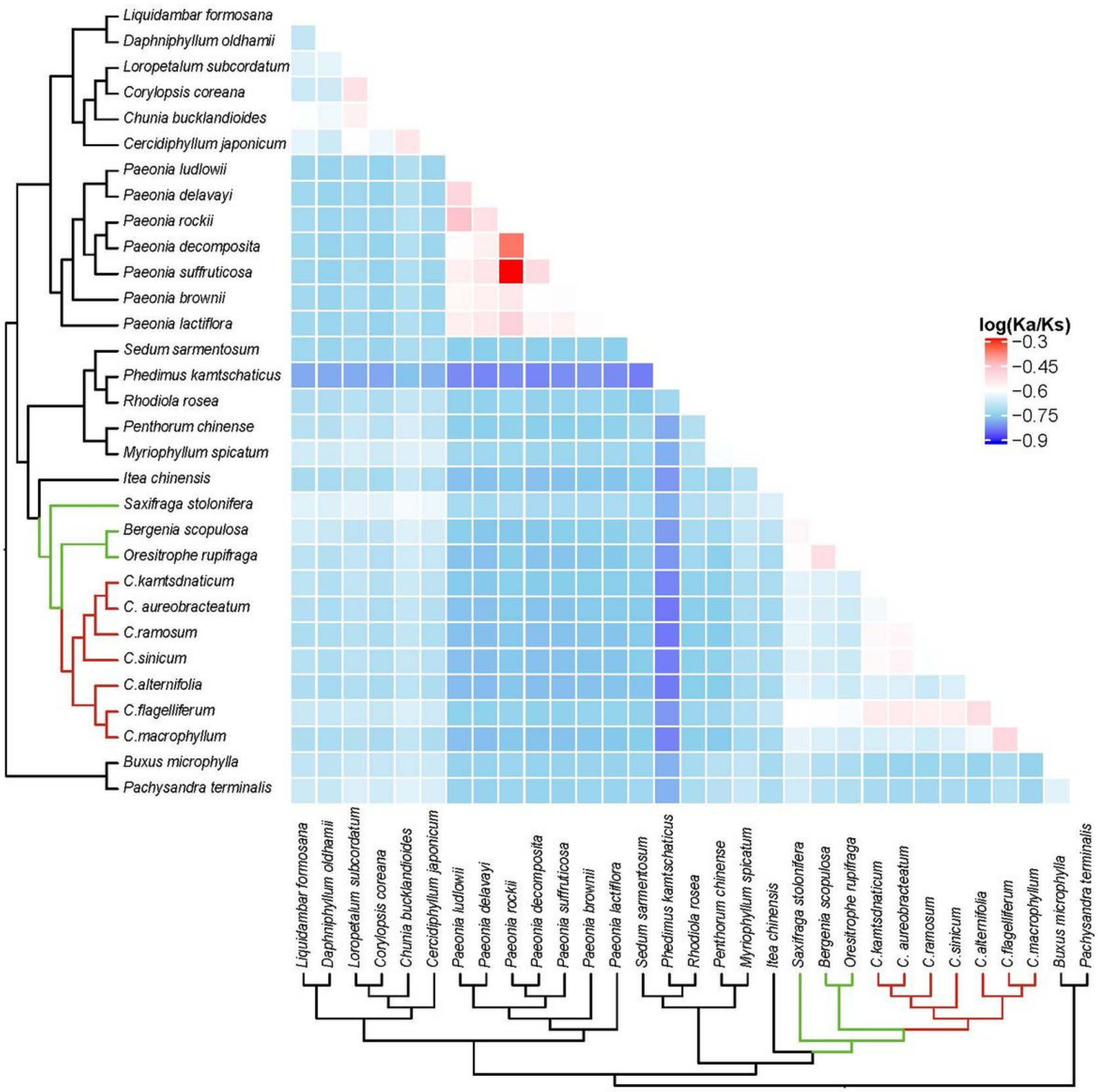

Figure 5

Pairwise Ka/Ks ratios in Saxifragaceae and other families. This heatmap shows pairwise Ka/Ks ratios between every sequence in the multigene nucleotide alignment. Chrysosplenium is shown on red branches. The scale factors associated with each value are shown on the right-hand side of the figure

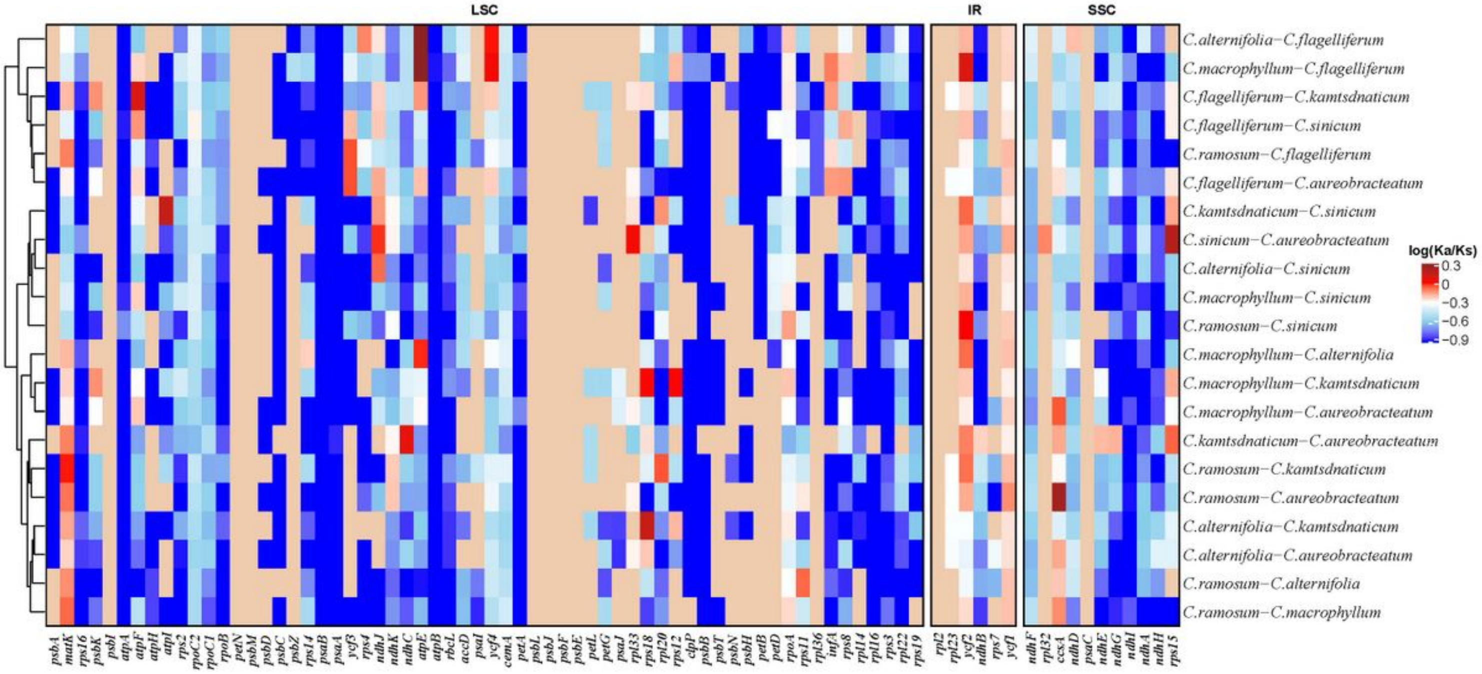

Figure 6

Pairwise Ka/Ks ratios in Chrysosplenium in different genes. This heatmap shows pairwise Ka/Ks ratios among each individual gene in the Chrysosplenium species.

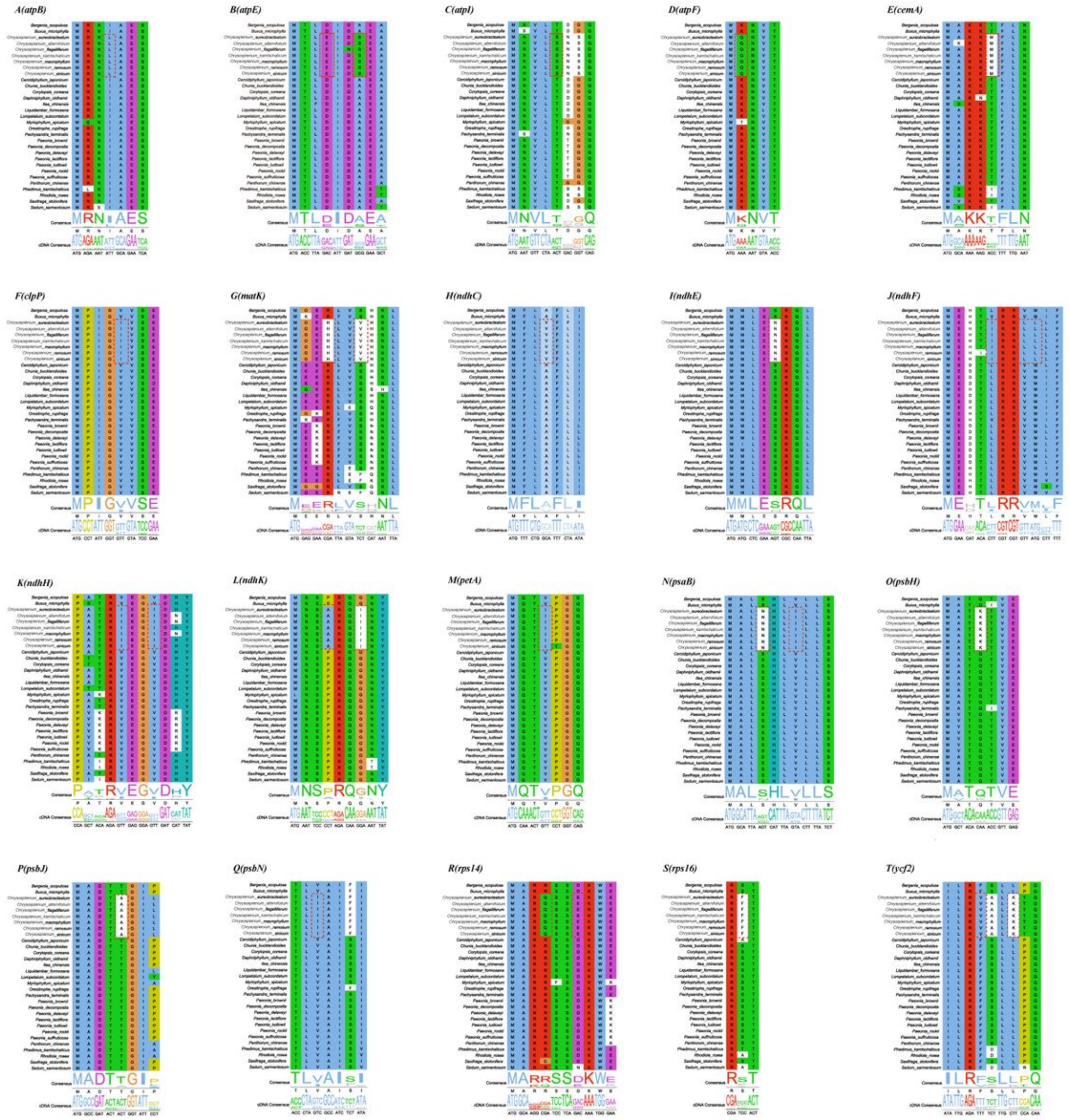

Figure 7

The partial alignment of 19 genes suggesting sites with positive selection in the BEB test. The red blocks stand for the amino acids in Chrysosplenium with a high BEB posterior probability.

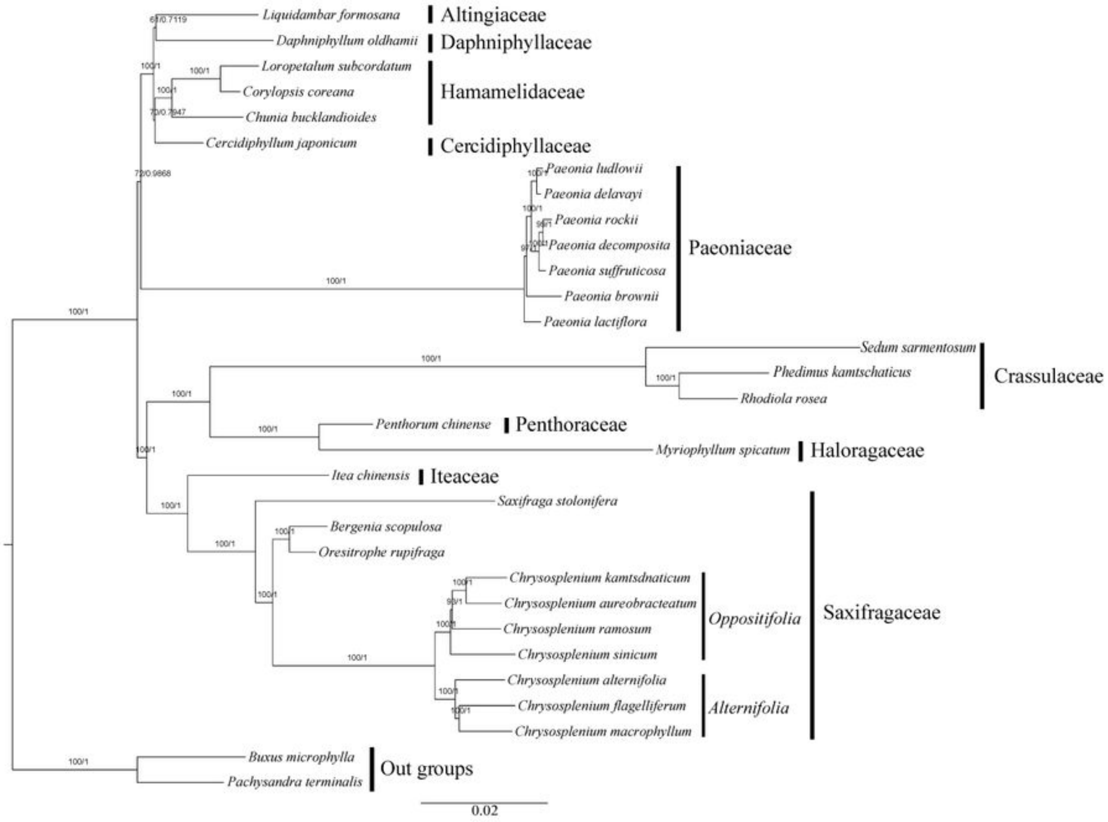

Figure 8

Phylogenetic tree reconstructed by Maximum likelihood (ML) and Bayesian inference (BI) analysis based on the whole chloroplast protein-coding genes of 31 species. The ML topology is indicated with ML bootstrap support values and BI posterior probabilities at each node.

Supplementary Files

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

- SupplementaryFile1.docx
- SupplementaryFile2.xlsx
- SupplementaryFigureS1.pdf
- SupplementaryFigureS3.pdf
- SupplementaryFigureS2.pdf