Complete Chloroplast Genome of Calligonum Mongolicum: Genome Organization, Codon Usage Pattern, Phylogenetic Relationships, Comparative Structure and Adaptive Evolution Analysis

Huirong Duan
Chinese Academy of Agricultural Sciences Lanzhou Institute of Husbandry and Pharmaceutical Sciences
Qian Zhang
Chinese Academy of Agricultural Sciences Lanzhou Institute of Husbandry and Pharmaceutical Sciences
Hongshan Yang
Chinese Academy of Agricultural Sciences Lanzhou Institute of Husbandry and Pharmaceutical Sciences
Fuping Tian
Chinese Academy of Agricultural Sciences Lanzhou Institute of Husbandry and Pharmaceutical Sciences
Yu Hu
Chinese Academy of Agricultural Sciences Lanzhou Institute of Husbandry and Pharmaceutical Sciences
Chunmei Wang
Chinese Academy of Agricultural Sciences Lanzhou Institute of Husbandry and Pharmaceutical Sciences
Yuan Lu
Chinese Academy of Agricultural Sciences Lanzhou Institute of Husbandry and Pharmaceutical Sciences
Huijun Yuan
Lanzhou University of Technology
Guangxin Cui (✉ cuigx03@sina.com)
Chinese Academy of Agricultural Sciences Lanzhou Institute of Husbandry and Pharmaceutical Sciences

Research article

Keywords: Calligonum mongolicum, Chloroplast genome, Synonymous codon usage, Genomic structure, Phylogeny

Posted Date: September 1st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-49271/v1

Abstract

Background The perennial shrub of *Calligonum mongolicum* is a dominant native plant in all *Calligonum* species, which has the largest and most widespread geographic distribution in arid deserts of northern China. Understanding the phylogenetic relationship between *C. mongolicum* and closely related plant species will offer guidance on the classification and identification of inter-species and their varieties. The chloroplast (cp) genome is an optimal model to decipher phylogenetic relationships and genome evolution in related plant families. In the present study, the complete cp genome of *C. mongolicum* was sequenced, and the characteristics were described, then the genomic structure was compared to other three Polygonaceae species.

Results The cp genome of *C. mongolicum* was 162,124 bp in length with a quadripartite structure. A total of 131 functional genes were annotated, 14 different genes of which harbored introns and exons, 50 long repeat sequences and 244 simple sequences repeats were identified. Synonymous codon usage (SCU) analysis exhibited A/T preference, and 7 codons were identified as the optimal codons. Multivariate statistical analysis of parity rule 2, ENC-plot, and neutrality plot were combined conducted to imply natural selection as the crucial constraint in SCU bias in *C. mongolicum* cp genome.The phylogenetic tree showed that *Rumex acetosa* was the most related plant to *C. mongolicum*. From the comparative analysis of genomic structures, the inverted repeat regions (IRa and IRb) were less divergent than other parts and coding regions was relatively conserved than non-coding regions. Compared to other species in the Polygonaceae, the borders of IRb/SSC and SSC/IRa in *C. mongolicum* changed greatly. Furthermore, adaptive evolution analysis of 75 orthologous protein-coding genes indicated that only the *psbK* gene was under positive selection, which might be crucial in the adaptive evolution of *C. mongolicum*.

Conclusions Our results comprehensively depicts the architecture of *C. mongolicum* cp genome, and will lay a vigorous foundation for further study on molecular marker selection, phylogenetic analysis, and population researches in *Calligonum* species.

Background

Calligonum L. belonging to the family Polygonaceae, contains approximately 35 species of perennial xerophytic shrubs or subshrubs, are successful pioneer plants for desert vegetation succession, and distribute in northern Africa, Southern Europe and Asia [1, 2]. There are 24 *Calligonum* species in China and almost of which possess significant ecological values [3]. The *Calligonum* species are divided into four groups according to the surface appendage of the achene [4, 5]. *Calligonum mongolicum*, belonging to Sect. Medusa Sosk. et Alexandr., has the largest and most widespread geographic distribution in arid deserts of northern China of all *Calligonum* species [6]. As a perennial shrub, *C. mongolicum* is native dominant in active sand dunes, exhibits strong resistance of drought, salt-alkali, high temperature, barren soil, wind erosion, also propagation strategies through sexual and asexual ways, thus it is often used for vegetation restoration in the desert [7, 8]. In recent years, the population of *C. mongolicum* exhibits a fast expansion tendency in mobile sand dunes of the Minqin Desert [9]. Actually, the surface appendages of

the achene from *C. mongolicum* and *C. roborovskii* have similar characteristics, and the geographical distributions of *C. mongolicum* and *C. roborovskii* are overlapped to some certain areas, as a result causing difficulty in identifying *C. mongolicum* [10]. Understanding the phylogenetic relationship between *C. mongolicum* and closely related plant species will offer guidance on the classification and identification of inter-species and their varieties.

Phylogenetic relationships are revolutionarily and frequently analyzed by genome sequencing, including the chloroplast genome (cp genome), the mitochondrion genome, and nucleus genome [11, 12]. The complete cp genome, which possesses many characteristics of small size, simple and highly conserved structure, single parentalin heritance, and haploid nature, is widely applied for species identification, phylogenic analysis, and adaptive evolution analysis [13]. Huang et al. [14] sequenced the complete cp genomes of five *Dilciptera* species, and compared the interspecies relationships of the five species. Xue et al. [12] also finished the cp genome comparison of three economic trees from *Prunus*, found the genomic structure difference, revealed the possible molecular markers, and concluded the phylogenetic evolution relationships. Until now, studies on the cp genome sequences have been conducted in many plant species, for example, watercress, yellow mustard, *Echinacanthus attenuatus*, and so on [15–17]. Thus, the cp genome is an optimal model to decipher phylogenetic relationships and genome evolution in related plants families.

Chloroplasts is a photosynthetic organelle in plant cells that plays crucial roles in the photosynthesis and crucial metabolites biosynthesis, for example amino acids, starch, fatty acids and pigments [18]. In general, the length of the cp genome ranges from 120–160 kb, the difference of which is mainly due to inverted repeated regions (IR) expansion, contraction or loss [19]. The cp genome encodes 63–209 unique genes, and is conserved with quadripartite organizations including a pair of IR regions (IRa and IRb), a large single copy (LSC) region and a small single copy (SSC) region [20]. For majority of higher plants, chloroplast is highly conservative in genes number, arrangement order and function [21]. However, gene losses and genome rearrangement can be found in some leguminous plants and conifer algae [22, 23]. After the first cp genome data of tobacco published in 1986, more and more plant cp genomes are completed sequencing, yet none of species from *Calligonum* L. has obtained the complete cp genomes data [24].

In the current study, the cp genome of *C. mongolicum* was constructed firstly by using Illumina sequencing and integrating a combination of *de novo* sequencing and reference-guided assembly. Then, the whole cp genome characteristics of *C. mongolicum* were described, and the synonymous codon usage (SCU) pattern, simple sequence repeats (SSRs) and long repeats were analyzed. Besides, we compared the cp genome of *C. mongolicum* with the published cp genomes of other three related species, including genome structure, IR contraction and expansion, and selective pressure events. This study may provide positive clues for adaptive evolution analysis of the *Calligonum* species.

Materials And Methods

Materials

The wild seedlings of *C. mongolicum* were transplanted from the Minqin desert to Lanzhou Scientific Observation and Experiment Field Station of the Ministry of Agriculture for Ecological System in the Loess Plateau Area (36°01'N,103°45'E, altitude 1700 m), Gansu, China in 15 May 2019. The station was belonged to Lanzhou Institute of Husbandry and Pharmaceutical Science, Chinese Academy of Agricultural Sciences, and we were approved to perform plant species planting and cultivating. *C. mongolicum* is a common wild plant in the Minqin desert, thus collecting or transplanting it for scientific research is not restricted in Ganu Province. Fresh leaf samples of *C. mongolicum* were collected in 22 July 2019. The sample collecting procedure was complied with our institutional guidelines. The voucher specimen was identified formally by expert on plant taxonomy from Gansu Grass Variety Committee and kept in Herbarium of Lanzhou Institute of Husbandry and Pharmaceutial Science (CYSLS-CmZhang20190722). Samples were quickly frozen in liquid nitrogen and conserved in -80℃ for the subsequent analysis.

DNA Extraction, Illumina Sequencing, Assembly and Annotation

Genomic DNA was isolated by the Plant Genomic DNA Rapid Extraction Kit (Biomed Gene Technology) with the modified CTAB method [25]. 1% Agarose gel electrophoresis and Qubit Fluorometer (Invitrogen) were used to check DNA integrity and quality. One library (350 bp) was constructed by using pure DNA with the NEBNext®Ultra™ DNA Library Prep Kit for Illumina®. The library was sequenced with an Illumina NovaSeq platform (Beganen Tech Solution CO., Ltd, Wuhan, China) and finally 150 bp paired-end reads were generated. The unknown reads and low-quality reads were filtered using SOAPnuke software (version: 1.3.0). Chloroplast-like reads were identified by BLAST (E-value $\leq 1^{e-5}$) with other related species. Finally, the chloroplast-like reads were assembled using NOVOPlasty (version: 32) with the parameter of k-mer 39 to form a circular genome. The sequences were annotated using GeSeq, mainly containing coding gene prediction and non-coding RNAs annotation (rRNA and tRNA) [26]. The circular genome map of *C. mongolicum* was drawn through the OGDRAWv1.2 program [27].

Synonymous Codon Usage Bias Analysis

A number of the codon usage indicators were performed via the program codon W version 1.3 (https://sourceforge.net/projects/codonw/), including the relative synonymous codon usage value (RSCU), the effective number of codons (ENC), G + C content of the gene (GC), the frequency of the nucleotides G + C at the 3rd position of synonymous codons (GC_{3s}), and the base compositions (A_{3s}, T_{3s}, G_{3s}, and C_{3s}) [28]. The RSCU value and ENC value were used together to describe codon usage patterns [29]. The G + C content at the 1st, 2nd, 3rd of codons (GC_1, GC_2, GC_3) and the average GC content of the 1st and 2nd (GC_{12}) were calculated by Cusp function from EMBOSS (http://imed.med.ucm.es/EMBOSS/) [30]. Synonymous codons with RSCU values > 1.3 were identified as high frequency codons. The optimal codon of the gene was speculated as the codon with both the highest RSCU value and the largest ΔRSCU

[31]. Parity rule 2 (PR2) plot mapping analysis was constructed to show the relationship of the values A_3 / (A_3 + T_3) and G_3 / (G_3 + C_3), and the data were distributed into four quadrants in a scatter diagram [32]. ENC-plot mapping analysis was performed to reflect the relationship of the ENC values against the GC_{3S} values [33]. Neutrality plot mapping analysis was used to analyze the relationship of the GC_{12} values and GC_3 values of all the genes [34].

Long repeat sequences and SSRs analysis

Long repeats including forward repeats and reverse repeats were analyzed by REPuter (http://bibiserv.techfak.uni-bielefeld.de/reputer) [30]. The hamming distance was set to 0 and the minimal repeat size was 20 bp. The SSRs were analyzed through the Perl script MISA (version: 1.0), and the SSRs parameters were defined as follows, the threshold of mononucleotide SSRs was ten repeats, the thresholds of dinucleotide and hexanucleotide SSRs were five repeats [35].

Phylogenetic Analysis

Phylogenetic tree with the bootstrap replicates set to 1000 was constructed by neighbor joining (NJ) analysis through TreesBeST (Version: 1.9.2, http://www.mybiosoftware.com/treebest-1-9-2-softwares-phylogenetic-trees.html). The cp genomes of *C. mongolicum* and other 36 species were used to investigate the evolution of *C. mongolicum*. The cp genomes information of the 36 plant species were downloaded from NCBI database.

Genome Structure Comparison

Based on the above results of the phylogenetic analysis, the complete cp genome of *C. mongolicum* was compared with other three closely related species of *Rumex acetosa* (NC_042390.1), *Rheum palmatum* (NC_027728.1), and *Fagopyrum esculentum* (NC_010776.1) using the mVista program with the shuffle-LAGAN mode [36]. The annotation of *C. mongolicum* was used as reference. The IRscope tool was used to visualize the genes on the boundaries of the junction sites of these four closely related species [37].

Selective pressure analysis

The orthologous genes of the Polygonaceae family were identified by OrthoMCL [38]. The sequences alignment of each orthologous gene was conducted using MAFFT [39]. The non-synonymous substitution rate (K_A) and synonymous substitution rate (K_S) were calculated by PAML [40]. The ω value was the ratio of K_A/ K_S.

Results

Features of C. mongolicum Cp Genome

The cp genome of C. mongolicum was 162,124 bp in length, was comprised by a pair of IR regions (IRa and IRb) (30,512 bp), a large single copy (LSC) region of 87,718 bp and a small single copy (SSC) region

of 13,382 bp (Fig. 1). The nucleotide composition of C. mongolicum cp genome was enriched in A/T nucleotides. The A + T content of the cp genome were 62.5%, which was significantly higher than the overall G + C content. The A + T content of the IR regions were 58.66%, obviously lower than LSC and SSC regions (64.42% and 67.51%, respectively) (Table 1). Weak base composition asymmetry (A-T, C-G) was found in C. mongolicum cp genome.

The positions of the 131 functional genes annotated in C. mongolicum cp genome were shown in Fig. 1. 78 genes were protein-coding genes, accounting for the half portion (59.5%) of the total genes. The remaining genes included 45 tRNA genes and 8 rRNA genes. According to the different functions, all the annotated genes were classified into four classes, including photosynthesis, self-replication, biosynthesis, and unknown functions (Table S1). Seventeen genes were duplicated in the IR regions, harboring 6 protein coding genes, 4 rRNA genes, and 6 tRNA genes.

In C. mongolicum cp genome, 12 different genes possessed a single intron and two exons, containing 5 tRNA genes and 7 protein coding genes, whereas the protein coding gene of ycf3 and clpP had two introns and three exons (Table 2). Of the total intron-containing genes, the gene of trnK-UUU had the largest intron (2511 bp), and the trnL-UAA had the smallest intron (520 bp).

Synonymous Codon Usage Analysis

A total of 51 coding sequences (CDSs) with length longer than 300 bp were screened for synonymous codon usage (SCU) bias analysis. In general, the four nucleotides were unevenly represented in the 51 CDSs. Adenine (A) and thymine (T) were the most represented (43.3% and 46.4%, respectively), cytosine (C) and guanine (G) were the least represented (16.7% and 16.9%, respectively), The average GC content of the CDSs was 38.7%.We identified the total of 61 synonymous codons except for stop codons, among which, the total of 18 codons with RSCU value more than 1.3 was identified as high frequency synonymous codons, 29 codons with ΔRSCU value more than 0.08 were identified as the high expressed codons (Table 3 and Table S2). 7 codons with high frequency as well as high expression including TTT, GGA, CAT, AAA, TTA, AAT and CCT were identified as the optimal codons.

To further analyze the SCU pattern in C. mongolicum cp genome, multivariate statistical analysis of PR2, ENC-plot analysis, and neutrality plot were combined conducted. PR2 plot mapping showed that the genes distributed unevenly in the four quadrants centered on 0.5, most points located under the horizontal centered line of 0.5 (the ratio of $A_3/ (A_3 + T_3) < 0.5$) (Fig. 2a). ENC plot was used to analyze the codon usage variation of the 51 CDSs (Fig. 2b). A majority of the points were lying away from the expected curve, accompanied with a relative concentrate distribution, and except for some points (rp116, ycf2, ycf3, and so on) located on the curve. Besides, we performed neutrality plot analysis to reveal the relationship of GC_{12} and GC_3 (Fig. 2c). Only one gene of ycf2 located on the effected curve, the remaining genes were up the standard curve.

Long-Repeat Sequences and Simple Sequences Repeats (SSRs) Analysis

The long repeat sequences in C. mongolicum cp genome were searched by REPuter software. A total of 50 long repeats were detected, 44 were forward and 6 were reverse repeats (Table 4). A majority long repeat sequences were only located in intergenic spaces (IGSs) (47%), 39% long repeat sequences were distributed in different genes, and the remaining long repeats (14%) were detected both in IGSs and genes. It was worth noting that the six reverse repeats were all located in IGSs. Besides, a total of 17, 1, 12, and 10 repeats harbored only one region of LSC, SSC, IRa and IRb regions, respectively. Another 10 repeats were detected simultaneously in two regions. Ycf1 CDS possessed the highest number of long repeats (14) and the longest repeats at 45 bp.

A total of 244 SSRs were found in C. mongolicum cp genome using MISA perl script. Among the identified SSRs, 67.2% was located in the LSC regions, 23.0% and 9.8% were found in the IR and SSC regions, respectively (Fig. 3a). 158 SSRs were located in IGSs, 80 SSRs were found in the coding regions and only 6 were found in introns (Fig. 3b). The numbers of mono-, di-, tri-, and tetranucleotides were 147, 43, 4, and 7, respectively (Fig. 3c). Mononucleotide repeats were the most frequented, accounting for 60.2% of the total repeats, while dinucleotides repeats accounted for 17.6%, and other SSRs were less common. Among all the identified SSRs, 20 SSRs belonged to G/C types, and the remaining SSRs belonged to the A/T types.

Phylogenetic Analysis

The phylogenetic tree was constructed based on a multiple alignment of nucleotide sequences of complete cp genomes from 37 plant species (Fig. 4). Drosera rotundifolia was used as the outgroup. The results showed that the species in the Polygonaceae family formed a clade, C. mongolicum was clustered closely to R. acetosa, R. palmatum, Oxyria sinensis, and F. esculentum. Furthermore, R. acetosa was the most related plant to C. mongolicum.

Comparative Analysis of Genomic Structure

The cp genome of *C. mongolicum* was compared to its closely related species including *R. acetosa*, *R. palmatum*, and *F. esculentum* (Table 5). *C. mongolicum* had the largest cp genome size, the largest SSC region and the most tRNA genes. *F. esculentum* had the smallest cp genome size and the largest LSC region. To further verify the genome divergence among these four species, sequence identity was compared using mVISTA with *C. mongolicum* as a reference (Fig. 5). Generally, IR regions were relatively conserved, while LSC and SSC regions were more divergent. Higher divergence of conserved non-coding regions were found than coding regions, for example, the IGS regions of *rps16* and *tmQ-UUG*, *ycf3* and *tmS-GGA*. Besides, significant differences were found in the regions of coding genes (*petD* and *ndhA*) and non-coding RNAs (*tml-GAU*).

IR Contraction and Expansion

The LSC/IR and SSC/IR boundaries of the cp genomes of *C. mongolicum* and other three related plant species were compared (Fig. 6). Six different genes were located at the juncture of the LSC/IRb (*rps19*

and *rp12*), IRb/SSC (*ndhF*), SSC/IRa (*rps15* and *ycf1*), and IRa/LSC borders (*rp12* and *trnH*), respectively. The *ndhF* gene crossed the IRb/SSC border, with 62-95 bp lengths within IRb region. Compared to other species in the Polygonaceae, the borders of the IRb/SSC and SSC/Ira in *C. mongolicum* changed greatly. The LSC/IRb and IRa/LSC borders were relatively conserved in *C. mongolicum*, *R. palmatum*, and *F. esculentum*, however the *rps19* gene at the LSC/IRb border and the *trnH* gene at the IRa/LSC border in *R. acetosa* varied from the other three species.

Selective pressure events

A total of 75 orthologous protein-coding genes were found in the family of Polygonaceae. The ω values of most genes were lower than 1, except for the psbK gene found in the LSC region, which had a ω value of 1.0556 (Figure 7). The ω values of some genes were 0, such as psbI, petN, ycf3, psbE, petG, rps12, and ndhE.

Discussion

Features of the C. mongolicum Cp Genome

Cp genomes of land plants are mostly conserved in structures, gene content, and organization of content [41]. Generally, the cp genome is a typical quadripartite circular structure and composed by two IR regions, large (LSC) and small single-copy (SSC) regions [42]. However, Tao et al. [21] reported alfalfa had a special cp genome structure with only one IR region. Besides, the linear cp genome is existed which is different from the typical plant cp genomes with a single circular molecule [43]. In this study, the complete cp genome of C. mongolicum revealed a typical circular and quadripartite structure, implying the relatively conserved cp genome in land plants. Despite the structures of cp genomes in different plant species are overall conserved, the size of which varies from 107 kb to218 kb [19]. The cp genome of C. mongolicum was ~162 kb, longer than the closely related plant species of R. acetosa (~160 kb), R. palmatum (~161 kb), and F. esculentum(~159 kb) to a certain extent. We also found that the A + T contents of the IR regions were significantly lower than other regions, which was similar to the observations in Prunus species, Quercus acutissima, and Phleum pretense and so on [14, 42, 44].

In land plant cp genomes, gene and intron content are highly conserved, although losses of them have been found in many angiosperms [11]. Funk et al. [45] found the losses of ndh genes in Cuscuta reflexa, they speculated the genes might be transferred to nuclear or the genes did not part in the critical life development. Here, we analyzed the genes and intron contents in C. mongolicum cp genome. The cp genome of C. mongolicum exhibited a complete set of genes (131), suggesting these genes might be critical to its development. Our results were similar to the finding of 130 genes in Nelumbo nucifera cp genome [46]. The cp genomes of the earliest diverging angiosperms contain the complete repertoire of 18 genes with introns [11]. Additionally, in many plants, the loss phenomenon of introns within protein-coding genes is often occurred, for example, Cicer arietinum, Mahonia bealei, and Hordeum vulgare [47-49]. The proteins encoded by genes with intron loss possess diverse functions. In C. mongolicum cp genome, we searched a total of 14 different genes with intron, and 4 intron losses were found, including

ycf1, rpoC2, rps19, and ndhF. The genes with intron losses might endow C. mongolicum diverse functions on RNA polymerase, ribosomal proteins, and NADH dehydrogenase.

Synonymous Codon Usage Pattern

SCU bias reflects uneven usage of synonymous codons with the same amino acid, which is different among different species and genes [50]. The possible causes of SCU bias have been investigated in the genomes of numerous living organisms, for example, *Zea mays*, cotton, Arabidopsis and so on [51-53]. In this study, 51 CDSs of *C. mongolicum* cp genome were selected to analyze the SCU bias. AT/GC nucleotide usage differed among the three positions of codon, and the genes showed a preference for AT-ending codons, thus revealing a SCU imbalance of A/T and G/C at the third base position. This speculation was further confirmed by PR2 analysis. The similar observations were also found in *Elaeagnus angustifolia*, *Porphyra umbilicalis* and so on [17, 54].

In the case of random mutation or mutation pressure in a certain direction, there should be no change in the three different positions of each codon and the base content should be similar [31]. Thus, the preference for A/T ending bases would drive the observation of natural selection competing against mutation pressure. In order to analyze the two major evolutionary factors on codon usage in *C. mongolicum*, we constructed ENC plot analysis and neutrality plot analysis. ENC plot analysis reflects the relationship of ENC value and GC_{3s}, thus detecting the SCU variation among the genes [33]. The distribution comparison of genes and the standard curve could be indicative for some other factors except mutation pressure [33]. In our study, it was observed that a few genes were lying on the curve, which definitely originated from the extreme mutation pressure. However, a majority of the points were lying well below the curve, suggesting that a majority of genes in *C. mongolicum* cp genome had other SCU bias, for example natural selection. This hypothesis was largely supported by neutrality plot mapping analysis. Neutrality plot analysis is useful to compare the impacts of selection constraints and mutation on SCU [55]. The low correlation between GC_{12} and GC_3 shows that the base composition of the 3 positions are different, and the GC content of the cp genome is highly conserved, indicating that natural selection is the most important determinant of codon usage patterns [56]. In the present neutral graph, no correlation was found between GC_3 and GC_{12}, indicating strong difference appeared, and natural selection would be crucial for SCU bias in *C. mongolicum* cp genome.

Molecular Markers

Long repeat sequences are associated with the sequence divergence and rearrangement of the cp genomes for illogical recombination and slipped-strand mismatch [57]. In C. mongolicum cp genome, more long repeats were found in LSC region than in IRs and SSC regions, verifying an uneven distribution phenomenon of long repeats in cp genomes. Many studies indicate that the sequence divergences are higher in the LSC and SSC regions than IR regions, also higher in IGSs than coding regions [58, 59]. Thus, the long repeats appeared in LSC and SSC regions, such as ycf3, ndhA, psaA and psaB, might help reducing sequence mutations in these regions. For the abundant variable sites, the gene of ycf1 was been

reported to be used in DNA barcodes [60]. Huang et al. [14] recommended ycf1 could be used as a candidate molecular marker for the Diciptera species. Similar to the finding in Dicliptera species, ycf1 harbored the highest number and the longest length of long repeats in C. mongolicum cp genome, thus we speculated ycf1 might be used as an important molecular marker for phylogenetic elucidation in Calligonum species.

SSRs, known as microsatellites, have many forms in the cp genome, and the variations in SSRs copy numbers are different between plant species [61, 62]. SSRs are usually used as potential genetic markers for plant population genetics, polymorphism investigations, and evolutionary research [16, 63]. Different to the nuclear genome, SSR variations in the cp genome exhibit outstanding characteristics in evolutionary study since it is sensitive to population genetic effects and exploration the maternal gene flow of populations [64]. In the present study, the SSRs identified in C. mongolicum cp genome were inclined to A/T types, which were similar to the observations in three Prunus species and Nasturtium officinale [12, 15], supporting the supposition that cp SSRs are generally composed by poly-adenine/-thymine (polyA/T) repeats [42]. Similar to long repeats, morenumbers of SSRs in LSC region than IRs and SSC region also implied an imbalanced repeats distribution in the cp genomes. Besides, the presence of the SSRs is indicators of important hotspots for genome recombination [14]. SSRs in C. mongolicum cp genome mainly located in IGSs which possessed high variable feature in the cp genome, suggesting that these regions could be treated as mutational hotspots. All the long repeats and SSRs identified in this study might furnish more venues for potential genetic markers in species identification and phylogenetic researches of Calligonum species.

Phylogenetic Analysis and Genomic Structure Comparison

Cp genome displays vast phylogenetic information, is usually used for phylogenetic reconstruction and population studies [65, 66]. In order to explore the phylogeny location of C. mongolicum and clearly elucidatethe genetic evolutionary relationships within Polygonaceae, we performed phylogenetic analysis based on the cp genome of C. mongolicum and other 36 cp genomes data of plant species. From the phylogenetic tree, all the 5 studied Polygonaceae species were clustered together, and C. mongolicum was accommodated as the neighbouring clade to the branch of R. acetosa accompanied with a 100% bootstrap value, fully confirming the genus Calligonum as a member of Polygonaceae. Thus, the constructed phylogenetic tree could be useful to confirm the phylogenetic position of C. mongolicum and further understanding the phylogeny relationships among more Polygonaceae species in the future.

Although cp genomes have highly conserved structures of most plants, four regions have varied genome sizes [49, 67]. The genomic structure comparison of four Polygonaceae species including C. mongolicum showed that they possessed different sizes in the four regions, mainly due to their different genus classifications. Among the four cp genomes of Polygonaceae species, more divergences emerged in LSC region, and higher divergence of conserved non-coding regions were found than coding regions. Our results were similar to the reports in other angiosperms, for example, Kaempferia galangal [68] and four Echinacanthus species [17], implying that less divergence in the IR regions and coding regions possibly

might cause copy corrections in the process of gene conversion. Besides, as the most conserved regions, frequent expansions and contractions at the boundaries of SSC/IR and LSC/IR can reflect the taxa relationships, thereby recognizing as evolutionary signals [69]. Compared to C. mongolicum cp genome, the IR regions of R. acetosa cp genome exhibited a slight contraction, and to the contrary, in F. esculentum and R. palmatum exhibited a slight expansion. Thus, the contractions and expansions at the two boundaries would contribute to the size variations of the four Polygonaceae species cp genomes.

Adaptive evolution analysis

Adaptability improvement of a species during evolutionary progress is reflected by adaptive evolution [14]. Adaptive evolution, driven by evolutionary factors (for instance natural selection), leads to pressures and diversity at different biological organization processes [70]. Makalowski and Boguski [71] reported that the ω value had been widely applied to identify the evolutionary dynamics and explore adaptive characteristics among species. The gene is under positive selection when the ω value is more than 1. Otherwise, the gene is negatively selected. The positively selected genes play important roles in diverse environment adaptation [17]. In the present study, the ω value of psbK gene was more than 1, suggesting psbK was under positive selection. Similar to our result, psbK was also detected under positive selection in karst topography [17]. Thus, we speculated the gene psbK could play an important role in the adaptation evolutionary process of Polygonaceae species to the diverse environment, and its unique function needs to be validated in further study.

Conclusions

In the study, the complete cp genome of *C. mongolicum* was firstly depicted comprehensively, including genome features, SCU bias, identification of long-repeat sequence and SSRs, and adaptation evolution analysis. The cp genome of *C. mongolicum* was a typical quadripartite structure and 131 functional genes were annotated, while 4 intron losses including *ycf1*, *rpoC2*, *rps19*, and *ndhF* were found. Codons in *C. mongolicum* cp genome presented A/T ending preference, possibly caused by major natural selection constraints. The phylogenetic analysis among 37 species revealed *C. mongolicum* was closely related to *R. acetosa*. Besides, comparative analysis of genomic structure of *C. mongolicum* and other three Polygonaceae species was conducted, revealing more divergence in LSC and SSC regions than IR regions, also more divergence in IGSs than coding regions, and these divergent regions could be treated as mutational hotspots. Furthermore, expansions and contractions at SSC/IRs and LSC/IRs junctions were also analyzed. A total of 50 long repeats and 244 SSRs were identified. The adaptation evolution analysis showed that only *psbK* gene was positively selected, thus *psbK* might play a crucial role in the adaptation of Polygonaceae species. In summary, our results would lay a vigorous foundation for further study on molecular marker exploration, phylogenetic signature, and population studies in *Calligonum* species.

Abbreviations

CDS: Coding sequences; LSC:Large single copy; SSC:Small single copy; IR:Inverted repeat; Cp:Chloroplast; A:Adenine; T:Thymine; C:Cytosine; G:Guanine; SCU:Synonymous codon usage; RSCU:Relative synonymous codon usage; ENC:Effective number of codons; IGS:Intergenic spaces; SSR:Simple sequence repeat; A_3:Adenine content at the third position of synonymous codons; T_3:Thymine content at the third position of synonymous codons; C_3:Cytosine content at the third position of synonymous codons; G_3:Guanine content at the third position of synonymous codons; GC_{12}:Average G + C content at the first and second positions of synonymous codons; GC_3:G + C content at the third positions of synonymous codons; GC_{3S}:G + C frequencies of at the third positions of synonymous codons; PR2:Parity rule 2.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publish

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was funded by the National Science Foundation of China (31700338), the Science and Technology Innovation Program of Lanzhou Institute of Husbandry and Pharmaceutical Science, Chinese Academy of Agricultural Sciences (CAAS-LMY-04), the Innovation Project of Chinese Academy of Agricultural Sciences (CAAS-XTCX2016011-02), the Central Public-interest Scientific Institution Basal Research Fund (1610322019012), and the Gansu Provincial Science and Technology Major Projects (19ZD2NA002). None of the funding bodies have any role in the design of the study or collection, analysis, and interpretation of data as well as in writing the manuscript.

Author Contributions

D. H. R. and C. G. X. designed the experiments. Z. Q. and Y. H. S. performed the experiments. T. F. P., H. Y. and W. C. M. analyzed the sequencing data. D. H. R. wrote the paper. L. Y., Y. H. J. and C. G. X. revised this paper. All authors have read and approved the manuscript.

Availability of data and materials

The raw sequence data of *C. mongolicum* cp genome have been submitted to GenBank under accession number MT568767. The address is as follows: https://www.ncbi.nlm.nih.gov/genbank.

References

1. Ren J, Tao L, Liu XM. Effect of sand burial depth on seed germination and seedling emergence of *Calligonum* L. species. J Arid Environ. 2002;51:603–11.
2. Ren J, Tao L. Effects of different pre-sowing seed treatments on germination of 10 *Calligonum* species. Forest Ecol Manag. 2004;195:291–300.
3. Mao ZM, Pan BR. The classification and distribution of *Calligonum* spp in China. Acta Phytotaxon Sin. 1986;24:98–107. (in Chinese).
4. Mao ZM. Flora of China. Vol. 25. Beijing: Beijing Science Press, China; 1998. pp. 120–33.
5. Wen ZB, Li Y, Zhang HX, Meng HH, Feng Y, Shi W. Species-level phylogeographical history of the endemic species *Calligonum roborovskii* and its close relatives in *Calligonum* section *Medusa* (Polygonaceae) in arid north-western China. Bot J Linn Soc. 2016;180:542–53.
6. Shi W, Pan BR, Gaskin JF, Kang XS. Morphological variation and chromosome studies in *Calligonum mongolicum* and *C. pumilum* (Polygonaceae) suggests the presence of only one species. Nord J Bot. 2009;27:81–5.
7. Fan BL, McHugh AD, Guo SJ, Ma QL, Zhang JH, Zhang XJ, et al. Factors influencing the natural regeneration of the pioneering shrub *Calligonum mongolicum* in sand dune stabilization plantations in arid deserts of northwest China. Ecol Evol. 2018;8:2975–84.
8. Zhang Q, Zhu XT. Microsatellite DNA loci from the drought desert plant *Calligonum mongolicum* Turcz. (Polygonaceae). Conserv Genet. 2009;10:1891–93.
9. Fan BL, Zhou YF, Ma QL, Yu QS, Zhao CM. Sun K. The bet-hedging strategies for seedling emergence of *Calligonum mongolicum* to adapt to the extreme desert environments in Northwestern China. Front Plant Sci. 2018;9:1167.
10. Liu N, Feng Y, Guan KY, Fan YQ, Chen JJ. Geographic distribution of *Calligonum mongolicum*. Arid Zone Res. 2015;32:753–59.
11. Jansen RK, Cai ZQ, Raubeson LA, Daniell H, dePamphilis CW, Leebens-Mack J, et al. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci USA. 2007;104:19369–74.
12. Xue S, Shi T, Luo WJ, Ni XP, Iqbal S, Ni ZJ, et al. Comparative analysis of the complete chloroplast genome among *Prunus mume*, *P. armeniaca*, and *P. salicina*. Hortic Res. 2019;6:89.
13. Raubeson LA, Peery R, Chumley TW, Dziubek C, Fourcade HM, Boore JL, et al. Comparative chloroplast genomics: analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. BMC Genom. 2007;8:174.
14. Huang SN, Ge XJ, Cano A, Salazar BG, Deng YF. Comparative analysis of chloroplast genomes for five *Dicliptera* species (Acanthaceae): molecular structure, phylogenetic relationships, and adaptive evolution. Peer J. 2020;8:e8450.
15. Yan C, Du J, Gao L, Li Y, Hou X. The complete chloroplast genome sequence of watercress (*Nasturtium officinale* R. Br.): genome organization, adaptive evolution and phylogenetic

relationships in Cardamineae. Gene. 2019;699:24–36.

16. Du XY, Zeng T, Feng Q, Hu LJ, Luo X, Weng QB, et al. The complete chloroplast genome sequence of yellow mustard (*Sinapisalba* L.) and its phylogenetic relationship to other Brassicaceae species. Gene. 2020;731:144340.
17. Gao C, Deng Y, Wang J. The complete chloroplast genomes of *Echinacanthus* species (Acanthaceae): phylogenetic relationships, adaptive evolution, and screening of molecular markers. Front Plant Sci. 2019;9:1989.
18. 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;76:273–97.
19. Daniell H, Lin CS, Yu M, Chang WJ. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biol. 2016;17:134.
20. Jansen RK, Raubeson LA, Boore JL, Depamphilis CW, Chumley TW, Haberle RC, et al. Methods for obtaining and analyzing whole chloroplast genome sequences. Method Enzymol. 2005;395:348.
21. Tao XL, Ma LC, Zhang ZS, Liu WX, Liu ZP. Characterization of the complete chloroplast genome of alfalfa (*Medicago sativa*) (Leguminosae). Gene Reports. 2017;6:67–73.
22. Wakasugi T, Tsudzuki J, Ito S, Nakashima K, Tsudzuki T, Sugiura M. Loss of all *ndh* genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*. Proc Natl Acad Sci USA. 1994;91:9794–98.
23. Kato T, Kaneko T, Sato S, Nakamura Y, Tabata S. Complete structure of the chloroplast genome of a legume, *Lotus japonicas*. DNA Res. 2000;7:323–30.
24. Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, et al. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J. 1986;5:2043–49.
25. Wang RJ, Cheng CL, Chang CC, Wu CL, Su TM, Chaw SM. Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. BMC Evol Biol. 2008;8:36.
26. Doyle JJ, Doyle JL. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull. 1987;19:11–5.
27. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, et al. GeSeq-versatile and accurate annotation of organelle genomes. Nucleic Acids Res. 2017;45:W6–11.
28. Lohse M, Drechsel O, Kahlau S, Bock R. 2013. Organellar Genome DRAW– a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Res. 2013; W575–81.
29. Zhang YY, Shi E, Yang ZP, Geng QF, Qiu YX, Wang ZS. Development and application of genomic resources in an endangered palaeoendemic tree, *Parrotia subaequalis* (Hamamelidaceae) from eastern China. Front Plant Sci. 2018;9:246.

30. Gupta SK, Bhattacharyya TK, Ghosh TC. SCU in *Lactococcus lactis*: mutational bias versus translational selection. J Biomol Struct Dyn. 2004;21:527–36.
31. Guan DL, Ma LB, Khan MS, Zhang XX, Xu SQ, Xie JY. Analysis of codon usage patterns in *Hirudinaria manillensis* reveals a preference for GC-ending codons caused by dominant selection constraints. BMC Genom. 2018;19:542.
32. Sueoka N. Near homogeneity of PR2-bias fingerprints in the human genome and their implications in phylogenetic analyses. J Mol Evol. 2001;53:469–76.
33. Wright F. The 'effective number of codons' used in a gene. Gene. 1990;87:23–9.
34. Wei L, He J, Jia X, Qi Q, Liang ZS, Zheng H, et al. Analysis of codon usage bias of mitochondrial genome in *Bombyx mori* and its relation to evolution. BMC Evol Biol. 2014;14:262.
35. Thiel T, Michalek W, Varshney R, Graner A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor Appl Genet. 2003;106:411–22.
36. Poliakov A, Foong J, Brudno M, Dubchak I. Genome VISTA-an integrated software package for whole-genome alignment and visualization. Bioinformatics. 2014;30:2654–55.
37. Amiryousefı A, Hyvönen J, Poczai P. IRscope: an online program to visualize the junction sites of chloroplast genomes. Bioinformatics. 2018;34:3030–31.
38. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13:2178–89.
39. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biol Evol. 2013;30:772–80.
40. Yang ZH. PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biol Evol. 2007;24:1586–91.
41. Wang XM, Zhou T, Bai GQ, Zhao YM. Complete chloroplast genome sequence of *Fagopyrum dibotrys*: genome features, comparative analysis and phylogenetic relationships. Sci Rep. 2018;8:12379.
42. Li X, Li YF, Zang MY, Li MZ, Fang YM. Complete chloroplast genome sequence and phylogenetic analysis of *Quercus acutissima*. Int J Mol Sci. 2018;19:2443.
43. Oldenburg DJ, Bendich AJ. The linear plastid chromosomes of maize: terminal sequences, structures, and implications for DNA replication. Curr Genet. 2016;62:431–42.
44. Cui GX, Wang CM, Wei XX, Yang HS, Lu Y, Wang XL, et al. Analysis of synonymous codon usage of the complete chloroplast genome in *Phleum pretense* cv. Minshan. Int J Agric Boil. 2020;24:352–58.
45. Funk H, Berg S, Krupinska K, Maier UG, Krause K. Complete DNA sequences of the plastid genomes of two parasitic flowering plant species, *Cuscuta reflexa* and *Cuscuta gronovii*. BMC Plant Bio. 2007;7:45.
46. Wu ZH, Gui ST, Quan ZW, Pan L, Wang SZ, Ke WD, et al. A precise chloroplast genome of *Nelumbo nucifera* (Nelumbonaceae) evaluated with Sanger, Illumina MiSeq, and PacBio RS II sequencing

platforms: insight into the plastid evolution of basal eudicots. BMC Plant Bio. 2014;14:289.

47. Jansen RK, Wojciechowski MF, Sanniyasi E, Lee SB, Daniell H. Complete plastid genome sequence of the chickpea (*Cicer arietinum*) and the phylogenetic distribution of *rps12* and *clpP* intron losses among legumes (Leguminosae). Mol Phylogenet Evol. 2008;48:1204–17.
48. Ma J, Yang BX, Zhu W, Sun LL, Tian JK, Wang XM. The complete chloroplast genome sequence of *Mahonia bealei* (Berberidaceae) reveals a significant expansion of the inverted repeat and phylogenetic relationship with other angiosperms. Gene. 2013;528:120–31.
49. Saski C, Lee SB, Fjellheim S, Guda C, Jansen RK, Luo H, et al. Complete chloroplast genome sequences of *Hordeum vulgare*, *Sorghum bicolor* and *Agrostis stolonifera*, and comparative analyses with other grass genomes. Theor Appl Genet. 2007;115:571–90.
50. Karumathil S, Raveendran NT, Ganesh D, Kumar NS, Nair RR, Dirisala VR. Evolution of SCU bias in West African and Central African strains of monkeypox virus. Evol Bioinform. 2018;14:1–22.
51. Liu HM, He R, Zhang HY, Huang YB, Tian ML, Zhang JJ. Analysis of SCU in *Zea mays*. Mol Biol Rep. 2010;37:677–84.
52. Wang LY, Xing HX, Yuan YC, Wang XL, Saeed M, Tao JC, et al. Genome-wide analysis of codon usage bias in four sequenced cotton species. Plos One. 2018;13:e0194372.
53. Qiu S, Zeng K, Slotte T, Wright S, Charlesworth D. Reduced efficacy of natural selection on codon usage bias in selfing Arabidopsis and Capsella species. Genome Biol Evol. 2011;3:868–80.
54. Li GL, Pan ZL, Gao SC, He YY, Xia QY, Yan J, et al. Analysis of synonymous codon usage of chloroplast genome in *Porphyra umbilicalis*. Genes Genom. 2019;41:1173–81.
55. Sueoka N. Directional mutation pressure and neutral molecular evolution. P Natl Acad Sci USA. 1988;85:2653–57.
56. Zhang DS, Hu P, Liu TG, Wang J, Jiang SW, Xu QH, et al. GC bias lead to increased small amino acids and random coils of proteins in coldwater fishes. BMC Genom. 2018;19:315.
57. Nie XJ, Lv SZ, Zhang YX, Du XH, Wang L, Biradar SS, et al. Complete chloroplast genome sequence of a major invasive species, crofton weed (*Ageratina adenophora*). PLoS One. 2012;7:e36869.
58. Wang CL, Ding MQ, Zou CY, Zhu XM, Tang Y, Zhou ML, et al. Comparative analysis of four buckwheat species based on morphology and complete chloroplast genome sequences. Sci Rep. 2017;7:6514.
59. Chen N, Sha LN, Dong ZZ, Tang C, Wang Y, Kang HY, et al. Complete structure and variation of the chloroplast genome of *Agropyron cristatum* (L.) Gaertn. Gene. 2018;640:86–96.
60. Shingo K, Jocelyn B, Minako H, Yoshino H, Maya O, Midori I, et al. Uncovering the protein translocon at the chloroplast inner envelope membrane. Science. 2013;339:571–74.
61. George B, Bhatt BS, Awasthi M, George B, Singh AK. Comparative analysis of microsatellites in chloroplast genomes of lower and higher plants. Curr Genet. 2015;61:665–77.
62. Ebert D, Peakall R. Chloroplast simple sequence repeats (cpSSRs): technical resources and recommendations for expandingcp SSR discovery and applications to a wide array of plant species. Mol Ecol Resour. 2009;9:673–90.

63. Yamane K, Lü N, Ohnishi O. Multiple origins and high genetic diversity of cultivated radish inferred from polymorphism in chloroplast simple sequence repeats. Breed Sci. 2009;59:55–65.
64. Lee SL, Nkongolo K, Park D, Choi IY, Choi AY, Kim NS. Characterization of chloroplast genomes of *Alnusrubra* and *Betulacordifolia*, and their use in phylogenetic analyses in Betulaceae. Genes Genom. 2019;41:305–16.
65. Lemieux C, Otis C, Turmel M. Ancestral chloroplast genome in *Mesostigma viride* reveals an early branch of green plant evolution. Nature. 2000;403:649–52.
66. Chang CC, Lin HC, Lin IP, Chow TY, Chen HH, Chen WH, et al. The chloroplast genome of *Phalaenopsis aphrodite* (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. Mol Biol Evol. 2005;23:279–91.
67. Choi KS, Park SJ. The complete chloroplast genome sequence of *Aster spathulifolius*, (Asteraceae); genomic features and relationship with Asteraceae. Gene. 2015;572:214–21.
68. Li D, Zhao CH, Liu X. Complete chloroplast genome sequences of *Kaempferia galangal* and *Kaempferia elegans*: molecular structures and comparative analysis. Molecules. 2018;24:474.
69. Wang RJ, Cheng CL, Chang CC, Wu CL, Su TM, Chaw SM. Dynamics and evolution of the inverted repeat-large single copy junctions in the chloroplast genomes of monocots. BMC Evol Biol. 2008;8:36.
70. Scottphillips TC, Laland KN, Shuker DM, Dickins TE, West ST. The niche construction perspective: a critical appraisal. Evolution. 2013;68:1231–43.
71. Makalowski W, Boguski MS. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. P Natl Acad Sci USA. 1998;95:9407–12.

Tables

Table 1. Bases composition of the *C. mongolicum* cp genome.

Region	A (%)	T (U) (%)	C (%)	G (%)	A+T (%)	G+C (%)
LSC	31.74	32.68	18.24	17.35	64.42	35.58
SSC	35.78	31.73	17.08	15.41	67.51	32.49
IR	29.33	29.33	20.67	20.67	58.66	41.34
Total	31.16	31.34	19.06	18.44	62.5	37.5

Table 2. The statistics of exons and introns in genes from C. mongolicum cp genome.

Gene	Location	Length				
		Exon I	Intron I	Exon II	Intron II	Exon III
trnK-UUU	LSC	35	2511	37		
rps16	LSC	231	858	36		
atpF	LSC	411	758	144		
rpoC1	LSC	1607	776	430		
ycf3	LSC	147	738	122	840	124
trnL-UAA	LSC	35	520	50		
trnV-UAC	LSC	37	593	36		
clpP	LSC	270	633	292	1143	71
rpl2_1	IRa	393	662	435		
ndhB_1	IRa	777	679	756		
rps12_1	IRa	25	533	232		
trnI-GAU_1	IRa	37	944	35		
trnA-UGC_1	IRa	38	803	36		
ndhA	SSC	537	1200	561		
trnA-UGC_2	IRb	36	803	38		
trnI-GAU_2	IRb	35	944	37		
rps12_2	IRb	232	533	25		
ndhB_2	IRb	756	679	777		
rpl2_2	IRb	435	662	393		

Table 3. Codon usage and high frequency used codons in C. mongolicum cp genome. The highest frequency used codons (RSCU values >1.3) are in bold. RSCU: relative synonymous codon usage.

Amino acid	Condon	Number	RSCU	Amino acid	Condon	Number	RSCU
Ala(A)	**GCT**	**497**	**1.72**	Asn(N)	**AAT**	**722**	**1.56**
	GCC	208	0.72		AAC	202	0.44
	GCA	324	1.12	Pro(P)	**CCT**	**350**	**1.6**
	GCG	130	0.45		CCC	167	0.76
Cys(C)	**TGT**	**160**	**1.45**		CCA	239	1.09
	TGC	61	0.55		CCG	120	0.55
Asp(D)	**GAT**	**654**	**1.57**	Gln(Q)	**CAA**	**552**	**1.49**
	GAC	181	0.43		CAG	187	0.51
Glu(E)	**GAA**	**832**	**1.48**	Arg(R)	CGT	298	1.43
	GAG	290	0.52		CGC	68	0.33
Phe(F)	**TTT**	**765**	**1.34**		CGA	310	1.48
	TTC	374	0.66		CGG	85	0.41
Gly(G)	GGT	441	1.23		**AGA**	**338**	**1.62**
	GGC	182	0.51		AGG	155	0.74
	GGA	**561**	**1.57**	Ser(S)	**TCT**	**415**	**1.68**
	GGG	248	0.69		TCC	232	0.94
His(H)	**CAT**	**380**	**1.53**		TCA	290	1.17
	CAC	118	0.47		TCG	134	0.54
Ile(I)	**ATT**	**875**	**1.52**		AGT	305	1.23
	ATC	323	0.56		AGC	106	0.43
	ATA	525	0.91	Thr(T)	**ACT**	**397**	**1.56**
Lys(K)	**AAA**	**818**	**1.51**		ACC	191	0.75
	AAG	263	0.49		ACA	316	1.24
Leu(L)	**TTA**	**696**	**1.92**		ACG	115	0.45
	TTG	453	1.25	Val(V)	**GTT**	**422**	**1.49**
	CTT	455	1.26		GTC	147	0.52
	CTC	133	0.37		GTA	414	1.46
	CTA	303	0.84		GTG	150	0.53

	CTG	135	0.37	Tyr(Y)	**TAT**	**609**	**1.6**
Met(M)	ATG	474	1		TAC	151	0.4
Trp(W)	TGG	383	1				

Table 4. Long repeat sequences in the C. mongolicum cp genome. F, forward; R, reverse; IGS, intergenic space.

ID	Repeat start 1	Type	Size (bp)	Repeat start 2	Mismatch(bp)	E-Value	Gene	Region
1	101960	F	42	128315	0	3.82E-16	IGS, *ndhA*	IRa, SSC,
2	114286	F	45	114298	-3	2.29E-12	*ycf1*	IRa
3	135499	F	45	135511	-3	2.29E-12	*ycf1*	IRb
4	70628	F	38	115967	-1	1.12E-11	IGS, *ycf1*	LSC, IRa
5	45870	F	39	101962	-2	1.63E-10	*ycf3*, IGS	SSC, IRa
6	45870	F	39	128317	-2	1.63E-10	*ycf3*, *ndhA*	LSC, SSC
7	89639	F	39	89682	-2	1.63E-10	IGS	IRa
8	89646	F	32	89689	0	4.01E-10	IGS	IRa
9	160121	F	32	160164	0	4.01E-10	IGS	IRb
10	114296	F	35	114308	-2	3.35E-08	*ycf1*	IRa
11	128532	F	28	128561	0	1.03E-07	*ndhA*	SSC
12	28584	R	34	28584	-2	1.26E-07	IGS	LSC
13	110672	F	34	110704	-2	1.26E-07	IGS	IRa
14	139104	F	34	139136	-2	1.26E-07	IGS	IRb
15	112364	F	31	112392	-1	1.49E-07	IGS	IRa
16	137419	F	31	137447	-1	1.49E-07	IGS	IRb
17	52182	F	27	52196	0	4.10E-07	IGS	LSC
18	29564	F	30	29580	-1	5.77E-07	IGS	LSC

19	113225	F	35	113246	-3	1.11E-06	*ycf1*	IRa
20	136561	F	35	136582	-3	1.11E-06	*ycf1*	IRb
21	40309	F	26	42524	0	1.64E-06	*psaB*, *psaA*	LSC
22	113217	F	34	113238	-3	4.05E-06	*ycf1*	IRa
23	136570	F	34	136591	-3	4.05E-06	*ycf1*	IRb
24	40007	F	28	42243	-1	8.62E-06	*psaB*, *psaA*	LSC
25	45882	F	27	101974	-1	3.32E-05	*ycf3*, IGS	LSC, IRa
26	45882	F	27	128329	-1	3.32E-05	*ycf3*, *ndhA*	LSC, SSC
27	4793	F	32	121628	-3	5.37E-05	IGS	LSC, SSC
28	8935	F	29	37672	-2	9.37E-05	*trnS-GCU*, *trnS-UGA*	LSC
29	114310	F	29	114322	-2	9.37E-05	*ycf1*	IRa
30	135491	F	29	135503	-2	9.37E-05	*ycf1*	IRb
31	29571	F	23	29587	0	1.05E-04	IGS	LSC
32	52185	F	26	52215	-1	1.28E-04	IGS	LSC
33	40858	F	31	43082	-3	1.95E-04	*psaB*, *psaA*	LSC
34	10845	F	25	38665	-1	4.92E-04	IGS, *trnG-UCC*	LSC
35	28587	F	25	34468	-1	4.92E-04	IGS	LSC
36	112307	F	25	112457	-1	4.92E-04	IGS	IRa
37	137360	F	25	137510	-1	4.92E-04	IGS	IRb
38	113157	F	30	136649	-3	7.03E-04	*ycf1*	IRa, IRb

39	113163	F	30	136655	-3	7.03E-04	*ycf1*	IRa, IRb
40	30761	R	27	30761	-2	1.30E-03	IGS	LSC
41	136577	F	27	136598	-2	1.30E-03	*ycf1*	IRb
42	38853	F	21	69758	0	1.68E-03	IGS, *trnP-UGG*	LSC
43	44430	R	21	50104	0	1.68E-03	IGS	LSC
44	59741	R	21	59741	0	1.68E-03	IGS	LSC
45	112374	F	21	112402	0	1.68E-03	IGS	IRa
46	139117	F	21	139149	0	1.68E-03	IGS	IRb
47	52199	F	24	52215	-1	1.89E-03	IGS	LSC
48	7567	R	29	7572	-3	2.53E-03	IGS	LSC
49	15150	R	29	121615	-3	2.53E-03	IGS	LSC, SSC
50	114278	F	29	114314	-3	2.53E-03	*ycf1*	IRa

Table 5. Summary of cp genome features of C. mongolicum and other three related plants species.

Genome features	*Calligonum mongolicum*	*Fagopyrum esculentum*	*Rheum palmatum*	*Rumex acetosa*
Genome Size (bp)	162,124	159,599	161,541	160,269
LSC length (bp)	87,718	84,884	86,518	86,135
SSC length (bp)	13,382	13,344	13,111	13,128
IR length (bp)	30,512	30,685	30,956	30,503
Number of genes	131	131	131	129
Number of protein coding genes	78	72	81	67
Number of tRNA genes	45	37	37	36
Number of rRNA genes	8	8	8	8

Figures

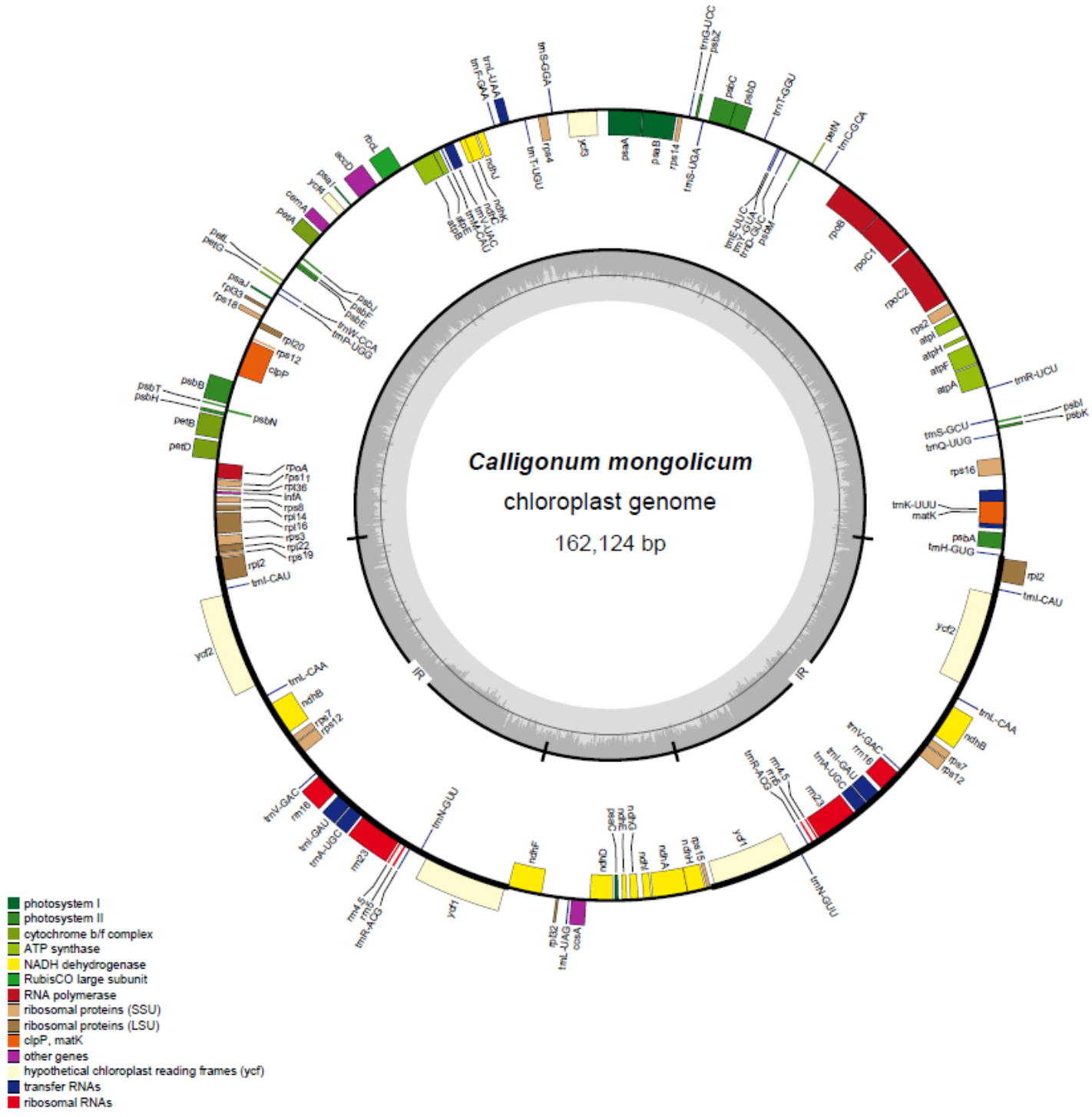

Figure 1

Cp genome map of C. mongolicum. Genes inside the circle and those outside are transcribed clockwise and counter-clockwise, respectively. Different colored flags show genes with different functions. The GC and AT content are represented by the darker gray and the lighter gray, respectively.

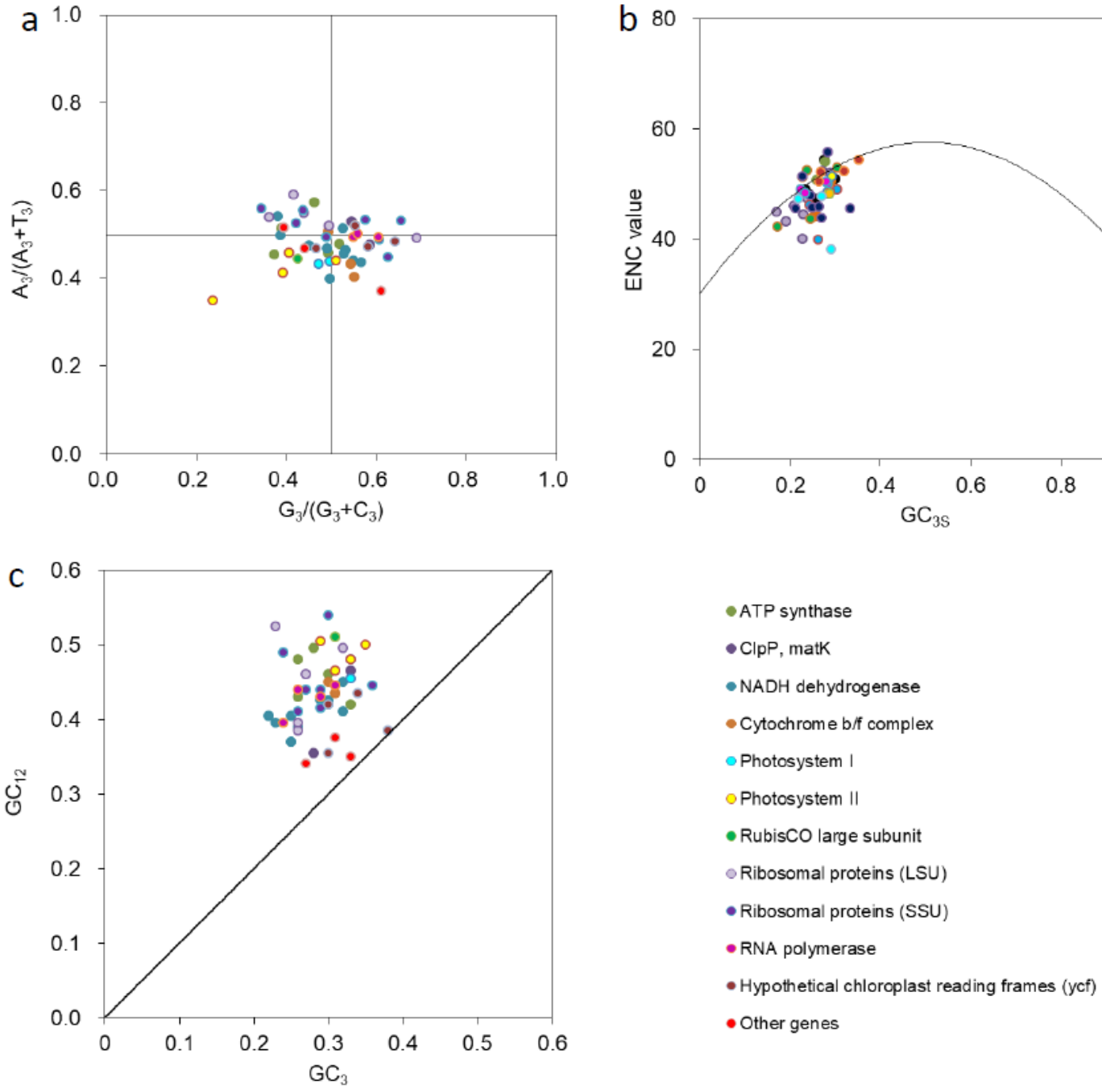

Figure 2

Multivariate statistical analysis for genes in C. mongolicum cp genome. a, PR2 analysis, A3, T3, C3 and G3 represents nucleotide A, T, C and G content at the third position of synonymous codons, respectively; b, ENC-plot analysis, ENC: effective number of codons. GC3s: the frequencies of nucleotide G + C at the third position of codons. The expected relationship between ENC values and GC3s was reflected on the curve with random codon usage assumption; c, Neutrality plot analysis. GC12: the average content of nucleotide G + C at the first and second positions of synonymous codons. GC3: the content of nucleotide

G + C at the third position of synonymous codons. The curve on the neutrality plot shows that the value of GC12 is equal to GC3.

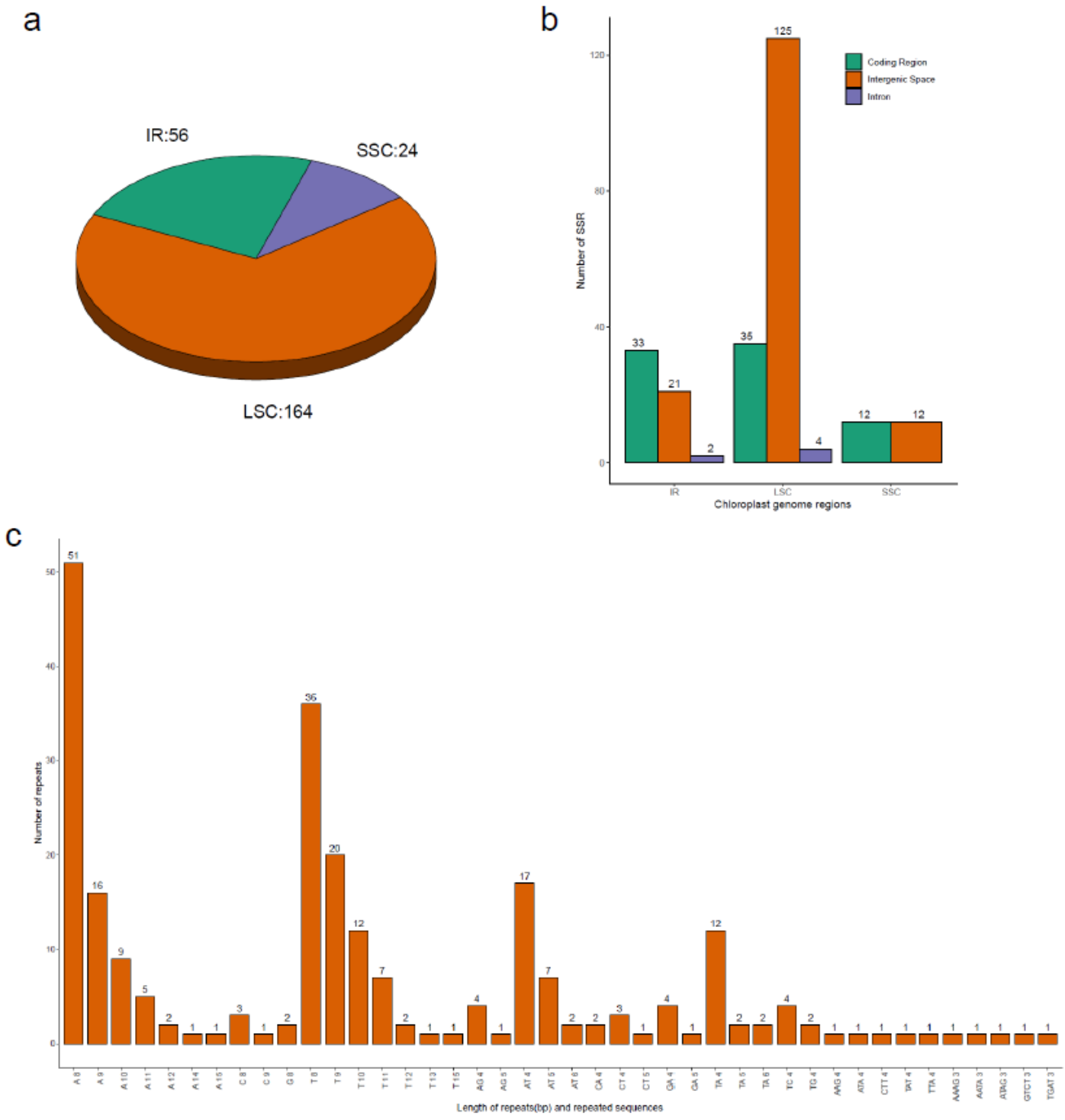

Figure 3

The simple sequence repeats (SSRs) in C. mongolicum cp genome. a, Distribution of SSRs in the LSC, SSC and IR regions; b, Presence of SSRs in the protein-coding regions, intergenic spaces (IGSs) and introns of LSC, SSC and IR regions; c, Types of nucleotides.

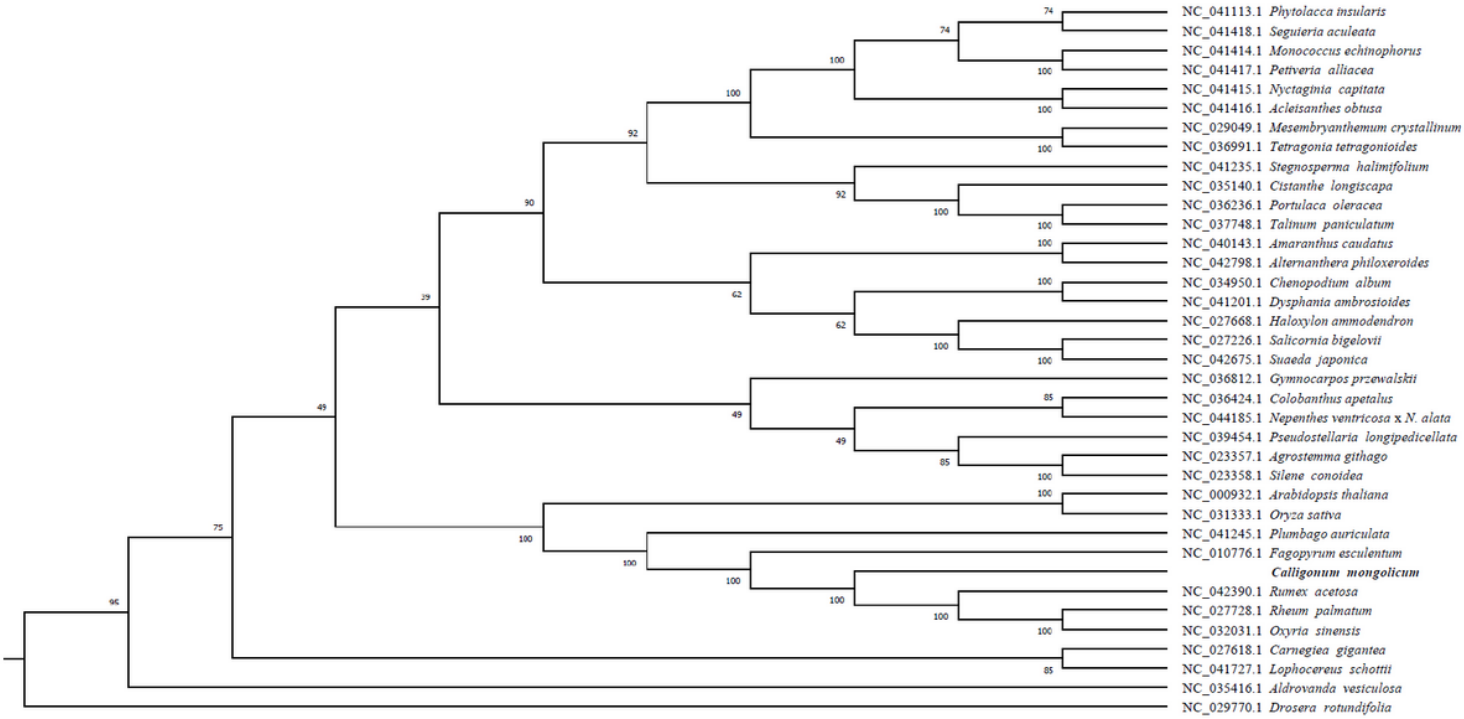

Figure 4

Phylogenetic relationship analysis of 37 species based on complete cp genomes. The phylogenetic tree was constructed by the neighbor-joining method. The bootstrap values were shown next to the branches based on 1000 replicates. The cp genome accession numbers from Genbank were labeled near to the names of plant species. C. mongolicum cp genome was in bold.

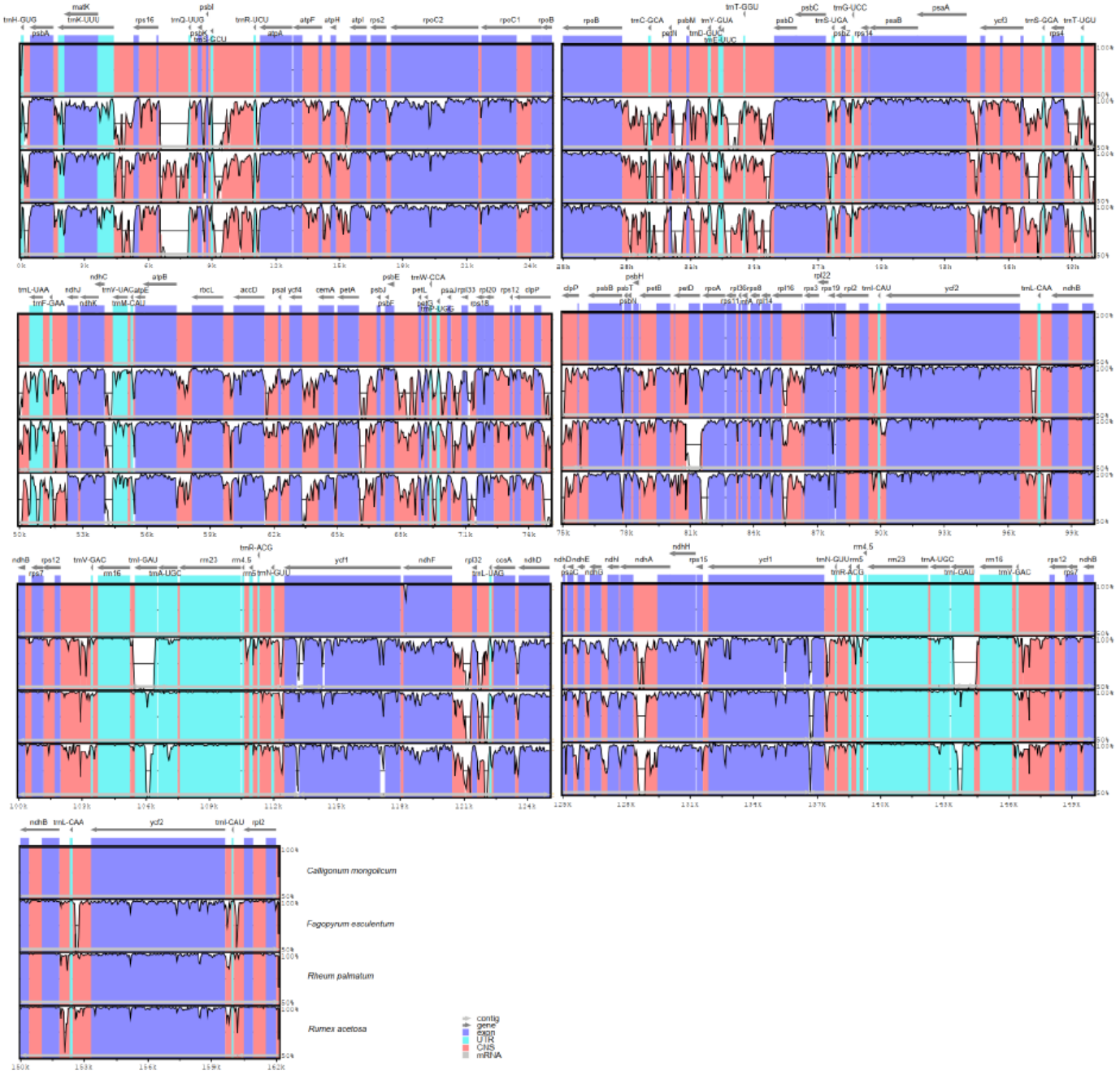

Figure 5

Complete cp genome comparison of C. mongolicum and other three related plants species. The cp genome of C. mongolicum was the reference and shown on the X-axis. The gene's orientations were shown by the grey arrows combined thick black lines on the top of the alignment. The Y-axis denotes identity ranging from 50% to 100%. UTR, untranslated regions. CNS, conserved non-coding sequences.

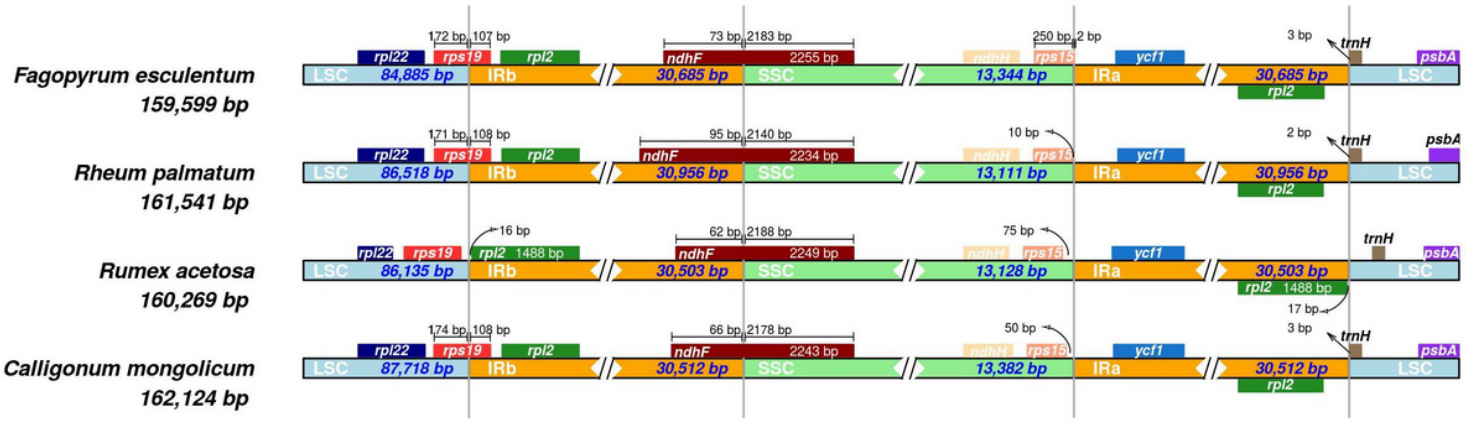

Figure 6

Comparison of junctions of the large single copy (LSC), small single copy (SSC), and inverted repeat (IR) in cp genomes of C. mongolicum and other three related plant species. Genes and regions are devided by colored boxes. The base lengths (bp) represent the gaps between the genes and the borders. Extensions are displayed above the boxes. The lengths of the four regions are also shown in the box.

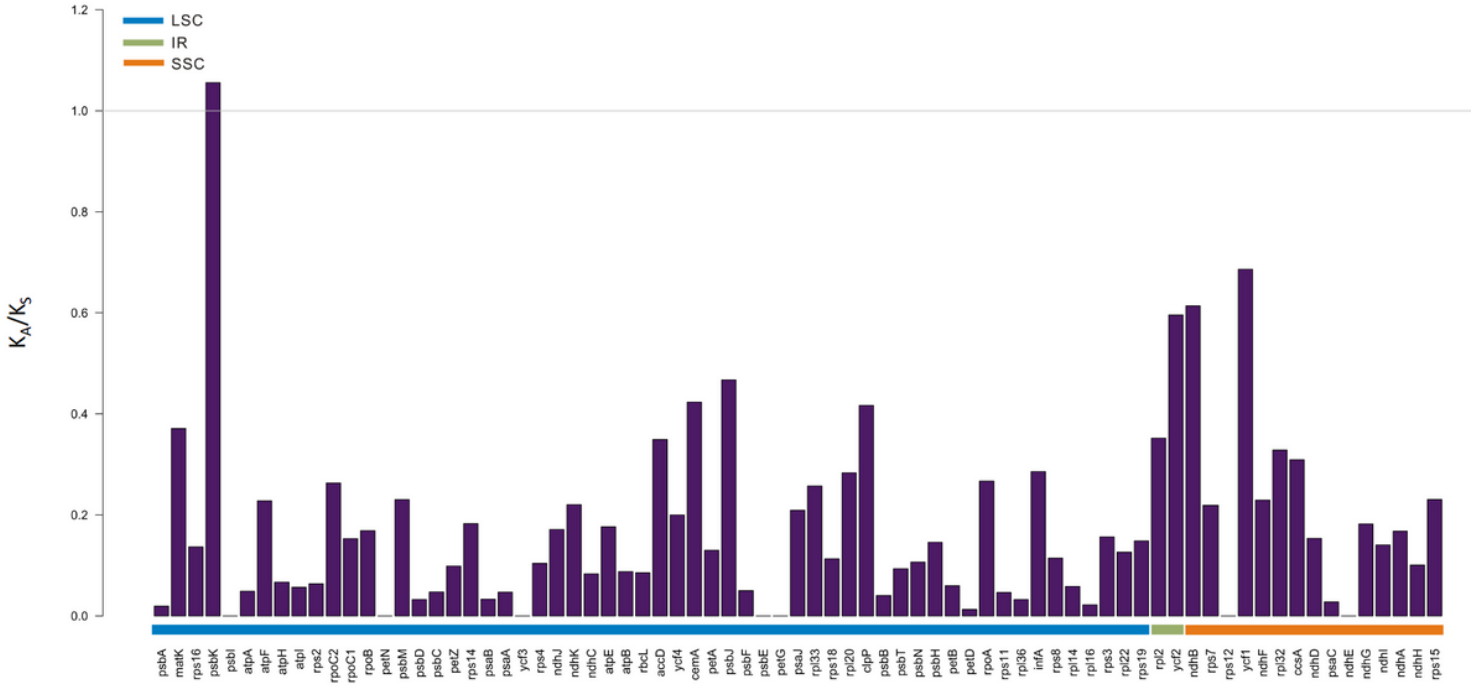

Figure 7

ω values among the five Polygonaceae species. ω value = KA/KS. KA, the nonsynonymous substitution rate. KS, the synonymous substitution rate. The region of gene distribution was shown in different color blocks.

Supplementary Files

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

- SupplementaryInformation.pdf