The chloroplasts genomic analyses of Caragana arborescens and Caragana opulens

LiE Liu
Qinghai University
HongYan Li
Qinghai University
JiaXin Li
Qinghai University
XinJuan Li
Qinghai University
Na Hu
Northwest Institute of Plateau Biology
Honglun Wang
Northwest Institute of Plateau Biology
Wu Zhou (zhouwu870624@qhu.edu.cn)
Qinghai University

Research Article

Keywords: Caragana arborescens, Caragana opulens; plastid genomes, comparative genomics, phylogenetic

Posted Date: June 16th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3036665/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Genomic Data on February 9th, 2024. See the published version at https://doi.org/10.1186/s12863-024-01202-4.

Abstract

Background

Numerous species within the genus *Caragana* have high ecological and medicinal value. In this genus, however, species identification based on morphological characteristics is quite complicated; this issue can be resolved by analyzing the complete plastid genomes.

Results

We obtained the chloroplast genomes of two species using Illumina sequencing technology: *Caragana arborescens* and *Caragana opulens*, with lengths of 129,473 bp and 132,815 bp, respectively. The absence of inverted repeat sequences in the two species allowed them to be ascribed to the inverted repeat-lacking clade (IRLC). They comprise a total of 111 distinct genes (4 rRNA, 31 tRNA, and 76 protein-coding genes). In addition, 16 genes containing introns were identified in two genomes, the majority of which contained a single intron. *C. arborescens* and *C. opulens* were found to contain 129 and 229 repeats, as well as 277 and 265 simple repeats, respectively. The codon usage bias analysis revealed that the two *Caragana* species exhibit similar codon usage patterns. *rpoC2-rps2*, *accD-cemA*, *rps18-clpP*, *rpoA-rpl36*, and *rpl2-rpl23* were identified as the five regions most likely to be mutated based on analysis of nucleotide diversity (Pi). Analysis of sequence divergence revealed that certain intergenic regions (*matK-rbcL*, *psbM-petN*, *atpA-psbI*, *petA-psbL*, *psbE-petL*, and *rps7-rps12*) are highly variable. Phylogenetic analysis showed that *C. arborescens* and *C. opulens* were related and clustered together as the other four *Caragana* species. And the genus *Astragalus* and *Caragana* were relatively closely related.

Conclusions

In our research, the chloroplast genomes of *C. arborescens* and *C. opulens* were sequenced and their genomic structural characteristics were compared. We have also confirmed that both plants lack IR regions, which resulted in unclear boundary analysis, and that two plants could be classified as IRLC. This study provides a foundation for future phylogenetic research and the development of molecular markers for *Caragana* plants.

Background

Approximately 100 species of the genus *Caragana*, which belongs to the subfamily Papilionoideae of the family Fabaceae, are primarily found in arid and semiarid regions of Asia and Europe. The majority of plants in this genus can withstand adverse environmental conditions, including sterile soil, drought, cold, high temperatures, strong winds, and insect and disease damage[1]. China was home to a total of 66 species, 32 of which were endemic. In China, they were primarily found in areas of high altitude and harsh environments, such as shady and semi-shady terrain in the northwest, southwest, northeast, and north[2]. *Caragana* is a deciduous undershrub with extensive adaptability and strong stress resistance, and it is known to resist wind and fix sand[3]. In addition, the majority of them can fix nitrogen via nodules, thereby enhancing soil fertility, preventing dust cyclones, and preventing land desertification[4]. *Caragana arborescens*, also known as

Siberian pea shrub, is typically found in Northeast China, North China, and Northwest China[5, 6]. It has a height of 4–5 meters. The species, which blooms in May with yellow flowers and pods that mature in midsummer, is typically used for foliage and garden decoration[7]. *Caragana opulens* is a shrub with a yellow corolla that thrives in the hills up to 3400 meters above sea level in North China, Northwest China, and Southwest China and is distributed throughout these regions[8]. In addition, previous research has demonstrated that numerous species of this genus possess outstanding pharmacological properties, including anti-cancer, anti-HIV, anti-rheumatoid arthritis, and hypertension[1, 9, 10]. *Caragana arborescens* has been documented in traditional Chinese medicine and is a significant Mongolian medicine used to treat pulmonary hemorrhage and rheumatism[1].

Current studies have revealed that the CP genome of *Caragana* plants contains only 14 reports, that the quantity of data available for analysis is extremely limited, that the phylogenetic relationship of *Caragana* plants is unclear, and that there are also questions regarding the classification of medicinal plants[2, 11–13]. Therefore, it is crucial to discover a precise and convenient method for identifying *Caragana* plants.

Researchers have improved their understanding of chloroplasts over the past decade, including their origin, structure, evolution, and genetic engineering[14–16]. The chloroplast (cp) is derived from the hypothesis of bacterial endosymbiosis in eukaryotes; it is the site of photosynthesis in plants and has its own DNA[17]. Cp possesses its own DNA (cpDNA) and genetic system, which exist as covalent double-stranded circular DNA[18, 19]. With the rapid advancement of sequencing technology, scientists have discovered that the chloroplast genome contains more effective molecular markers, which facilitates the precise identification of species. The chloroplast genome is optimal for molecular identification, phylogenetic, and species conservation research[20]. Unlike the nuclear genome, the mitochondrial genome is characterized by unisexual inheritance, a simple structure, and more gene copies[19, 21]. Typically, the chloroplast genome is maternally inherited in angiosperms[22]. Its structure is comparatively stable and consists of a large single copy (LSC) and a small single copy (SSC) region separated by two inverted repeats (IRs)[23]. It has been reported that the phenomenon of inverted repeat-lacking clade (IRLC) occurs in leguminous plants[24–28]. There have been reports of eight species of *Caragana* plants with IRLC[2, 12, 13, 29]. With the refinement of *Caragana* chloroplast genome data, *Caragana* will presumably represent a broad IRLC spectral system for scientific investigation. Moreover, the chloroplast genome sequence offers more secure information for the study of genetic relationships, phylogeny, and population genomics among closely related species[29–31].

In this study, the complete chloroplast genomes of *C. arborescens* and *C. opulens* were obtained using Illumina sequencing technology, and their structural properties and phylogenetic relationships were elucidated. The completion of this endeavor has enriched the chloroplast genomic database of *Caragana*, which is anticipated to serve as a foundation for systematic evolution research and the protection and utilization of *Caragana*'s germplasm.

Results

Chloroplast genome assembly and features

C. arborescens and *C. opulens* chloroplast genomes were sequenced using the Illumina Novaseq platform. According to the sequencing results, the chloroplast whole genome sequence was assembled at 129,473 (Fig. 1A)and 132,815 base pairs (Fig. 1B). Due to the loss of the IR region, neither of their chloroplast genomes have the typical tetrad structure of most angiosperm chloroplast genomes, and their length has been shortened accordingly. Nonetheless, their genetic structures are extremely comparable.

In the chloroplast genomes of *C. arborescens* and *C. opulens*, there were 111 unique genes, including 76 protein-coding genes, 31 tRNA genes, and 4 rRNA genes, and their respective GC contents were 34.30% and 34.71% (Table 1), indicating that the GC content between the two species was extremely similar. This paper compares and analyzes the chloroplast genome sequences of six species of *Caragana* plants lacking the IR region. According to the results, the total length of their sequences varied between 129,331 and 133,122 base pairs. Due to the absence of the IR region, the chloroplast genome length of *C. korshinskii* was the shortest, at only 129,311 bp, and that of *C. rosea* was the longest, at a total length of 133,122 bp. In addition, the number of genes in *C. arborescens* and *C. opulens* was one gene greater than that of other species (tRNA encoded by the *trnN-GUU* gene), whereas the number of protein-coding genes and rRNA genes was consistent among the six plants. In terms of gene content, the number of protein-coding genes was the highest among the six species, comprising approximately half of the full-length genome, followed by the number of tRNA genes, whose length was shorter than that of other genes. *C. rosea* has the highest GC content in its chloroplast genome, at 34.84 percent, followed by *C. kozlowii* (34.5 percent), and *C. microphylla*, which has the lowest GC content, at 34.2 percent. We also examined the variations in GC concentration between the three gene types. The GC concentration of rRNA was over 50%, which was high and stable, followed by tRNA, and the GC content of protein-coding genes was approximately 37%. In conclusion, the sequence length and gene number of the chloroplast genomes of the six *Caragana* species were generally consistent, and the average GC content of the species was approximately 34%, which suggests that the evolution of the *Caragana* genus was relatively conservative.

Table 1
Summary of complete chloroplast genomes for six *Caragana* species.

Plastome Characteristics		***Caragana arborescens***	***Caragana opulens***	***Caragana kozlowii***	***Caragana rosea***	***Caragana microphylla***	***Caragana korshinskii***
Protein Coding gennes	Length(bp)	66,222	66,333	66,234	66,243	66,231	66,231
	GC(%)	36.89	37.01	37.03	37.13	36.88	36.88
	Length(%)	51.15	50.0	50.45	49.76	50.94	51.21
	Number	76	76	76	76	76	76
tRNA	Length(bp)	2,379	2,370	2,285	2,359	2,370	2,379
	GC(%)	52.74	52.83	53.15	52.73	53.14	53.05
	Length(%)	1.83	1.80	1.74	1.77	1.82	1.83
	Number	31	31	30	30	30	30
rRNA	Length(bp)	4,522	4,520	4,521	4,537	4,520	4,520
	GC(%)	54.8	54.56	54.75	54.77	54.82	54.82
	Length(%)	3.49	3.40	3.44	3.4	3.48	3.49
	Number	4	4	4	4	4	4
Total	Length(bp)	129473	132815	131274	133122	130029	129331
	Number Of genes	111	111	110	110	110	110
	GC(%)	34.3	34.71	34.5	34.84	34.26	34.36

Comparable to other species, the chloroplast genomes of *C. arborescens* and *C. opulens* encode three categories of genes (Table 2). Self-replication was associated with 57 genes. 3 subunits (large, small, and DNA-dependent RNA polymerase), including ribosomal RNA genes, transporting RNA genes, and encoding chloroplast RNA polymerase; 44 photosynthesis-related genes; other genes and unknown genes. In the chloroplast genomes of *C. arborescens* and *C. opulens*, 16 genes with introns were detected, of which one gene, ycf3, had two introns, and the remaining 15 genes (*trnK-UUU*, *trnV-UAC*, *trnL-CAA*, *rpoC1*, *atpF*, *trnG-UCC*, *clpP*, *petB*, *petD*, *rpl16*, *rpl2*, *ndhB*, *trnI-GAU*, *trnA-UGC*, *ndhA*) had only one intron (Table 3). Among these 16 intron-containing genes, the intron lengths of the two genes were remarkably similar.

Table 2
Genes in the chloroplast genome of _Caragana_ species.

Category	Group of genes	Name of genes
Self-replication	Proteins of large ribosomal subunit	_rpl14_, _rpl16_*, _rpl2_*, _rpl20_, _rpl23_, _rpl32_, _rpl33_, _rpl36_
	Proteins of small ribosomal subunit	_rps11_, _rps12_*, _rps14_, _rps15_, _rps18_, _rps19_, _rps2_, _rps3_, _rps4_, _rps7_, _rps8_
	Subunits of RNA polymerase	_rpoA_, _rpoB_, _rpoC1_*, _rpoC2_
	Ribosomal RNAs	_rrn16_, _rrn23_, _rrn4.5_, _rrn5_
	Transfer RNAs	_trnA-UGC_*, _trnC-GCA_, _trnD-GUC_, _trnE-UUC_, _trnF-GAA_, _trnG-GCC_, _trnG-UCC_*, _trnH-GUG_, _trnI-CAU_, _trnI-GAU_*, _trnK-UUU_*, _trnL-CAA_*, _trnL-UAA_, _trnL-UAG_, _trnM-CAU_, _trnN-GUU_(2), _trnP-UGG_, _trnQ-UUG_, _trnR-ACG_, _trnR-UCU_, _trnS-GCU_, _trnS-GGA_, _trnS-UGA_, _trnT-GGU_, _trnT-UGU_, _trnV-GAC_, _trnV-UAC_*, _trnW-CCA_, _trnY-GUA_, _trnfM-CAU_
Photosynthesis	Subunits of photosystem I	_psaA_, _psaB_, _psaC_, _psaI_, _psaJ_
	Subunits of photosystem II	_psbA_, _psbB_, _psbC_, _psbD_, _psbE_, _psbF_, _psbH_, _psbI_, _psbJ_, _psbK_, _psbL_, _psbM_, _psbN_, _psbT_, _psbZ_
	Subunits of NADH dehydrogenase	_ndhA_*, _ndhB_*, _ndhC_, _ndhD_, _ndhE_, _ndhF_, _ndhG_, _ndhH_, _ndhI_, _ndhJ_, _ndhK_
	Subunits of cytochrome b/f complex	_petA_, _petB_*, _petD_*, _petG_, _petL_, _petN_
	Subunits of ATP synthase	_atpA_, _atpB_, _atpE_, _atpF_*, _atpH_, _atpI_
	Large subunit of rubisco	_rbcL_
Other genes	Maturase	_matK_
	Protease	_clpP_
	Envelope membrane protein	_cemA_
	Acetyl-CoA carboxylase	_accD_

Notes: Gene*:Gene with one introns; Gene**:Gene with two introns; Gene(2):Number of copies of multi-copy genes.

Category	Group of genes	Name of genes
	c-type cytochrome synthesis gene	*ccsA*
Unknown	Conserved hypothetical chloroplast ORF	*ycf1*, *ycf2*, *ycf3***, *ycf4*

Notes: Gene*:Gene with one introns; Gene**:Gene with two introns; Gene(2):Number of copies of multi-copy genes.

Table 3
The intron-containing genes and the length of exons and introns in the chloroplast genomes of two *Caragana* species

Species	Gene	Exon I(bp)	Intron I(bp)	Exon II(bp)	Intron II(bp)	Exon III(bp)
C.arborescens	trnK-UUU	37	2488	29		
	trnV-UAC	39	574	37		
	trnL-CAA	37	550	50		
	ycf3	126	713	228	877	153
	rpoC1	432	789	1623		
	atpF	168	660	411		
	trnG-UCC	23	682	49		
	clpP	368	701	229		
	petB	6	818	642		
	petD	8	717	475		
	rpl16	9	1054	399		
	rpl2	396	685	435		
	ndhB	723	685	762		
	trnI-GAU	38	951	35		
	trnA-UGC	38	812	35		
	ndhA	552	1170	540		
C.opulens	trnK-UUU	37	2485	29		
	trnV-UAC	39	574	37		
	trnL-CAA	37	534	50		
	ycf3	126	702	228	875	153
	rpoC1	432	790	1623		
	atpF	168	679	411		
	trnG-UCC	23	682	49		
	clpP	368	1159	223		
	petB	6	826	642		
	petD	8	720	475		
	rpl16	9	1108	399		

Species	Gene	Exon I(bp)	Intron I(bp)	Exon II(bp)	Intron II(bp)	Exon III(bp)
	rpl2	396	692	435		
	ndhB	723	685	762		
	trnI-GAU	38	953	35		
	trnA-UGC	38	810	35		
	ndhA	552	1169	540		

Analyses of repetitive sequences and SSRs

Repeat sites are important in genomic evolution, such as in structural rearrangement and size-based evolution [32, 33]. In this study, we identified the repetitive sequences in the chloroplast genomes of *C. arborescens* and *C. opulens* and analyzed their content. The results indicated that the chloroplast genome with a repeat length greater than or equal to 30 bp contained four categories of repeats: forward (F), palindromic (P), reverse (R), and complementary (C) repeats. In the two plants, 129 (length range: 30–249 bp) and 229 (length range: 30–472 bp) repeats, respectively, were identified (Table S1). The length range of 30–49 bp sequences had the highest frequency among all classes of repetitive sequences (former: 68.22%, latter: 52.40%).

Structural analysis of the repetitive sequences showed that *C. arborescens* was composed of 85 forward repeats (65.89%), 36 palindromic repeats (27.91%), 7 reverse repeats (5.43%), and 1 complementary repeat (0.78%) (Fig. 2A, Fig. 2C), while there are no complementary repeats in the repeat sequence of *C. opulens*, which consists of three repeat types, including 165 forward repeats (72.05%), 62 palindromic repeats (27.07%), and 2 reverse repeats (0.87%) (Fig. 2B, Fig. 2C). The majority of repeat sequences exist in the IGS region, and the majority of them are forward repeats.

Numerous simple sequence repeats (SSRs) are present in the chloroplast genome of plants. This form of sequence is transmitted from parents to offspring. It has a relatively basic structure and low variability. SSRs are therefore more efficient molecular markers [34]. Using the software MISA v1.0, we identified a total of 18 varieties in the two Caragana plants. The chloroplast genomes of *C. arborescens* and *C. opulens* contain 277 and 265 SSR loci, respectively (Table S2). The proportion of mononucleotide in the two *Caragana* plants with the highest concentration were 57.04 and 63.40 percent, respectively. While dinucleotide and trinucleotide repeat sequences comprised 7.58 and 29.24 percent of the former, tetranucleotide repeat sequences comprised the smallest proportion (6.14 percent). In the latter, the proportions of dinucleotide, trinucleotide, and tetranucleotide repeat sequences were 4.91 percent, 28.68 percent, and 2.64 percent, respectively, while pentanucleotide represented the smallest proportion, 0.38 percent.

C. arborescens has the longest SSR on the ycf1 gene of the chloroplast genome, which is a single nucleotide repeat sequence (A) with a length of 46 bp, whereas *C. opulens* has the longest SSR, which is a mononucleotide (T) with a length of 26 bp (Table S3). In addition, the distribution of SSRs in coding and noncoding regions was analyzed. Figure 3A displays that the number of SSRs in the protein-coding region was significantly lower than in the non-coding region. The majority of these SSRs were A/T single nucleotide repeats; 158 and 167 of the two *Caragana* species contained A/T, while only one contained C/G(Fig. 3B, Fig. 3C). Similarly, the majority of dinucleotide repeats consist of AT/AT, resulting in a deviation in base composition, which is consistent with the finding that the overall AT content of plastids is greater than the GC content[35].

Codon usage bias analysis

In the evolution of biology, plastids exhibit a prevalent codon usage bias. By analyzing codon usage bias, which may penetrate the phylogenetic relationship between bionts and the molecular phylogeny of genes[36], it is possible to study the origin, mutation model, and evolution of species. We have analyzed the codon distribution conditions in all protein-coding genes in these two plants. The 76 protein-coding gene sequences of the two *Caragana* species were used to generate 12,812 codons in total. Leucine (Leu) was the amino acid with the highest content, accounting for 10.58% and 10.65%, respectively, followed by codons encoding isoleucine (Ile) (9% and 8.89%), while cysteine (Cys) had the lowest abundance among the two plants (Table S4).

In the meantime, we also independently calculated the relative synonymous codon usage (RSCU) values, using which we determined the codon usage bias of the two plants' chloroplast genomes (Fig. 4). When the RSCU value is greater than one, the codon is considered optimal. Among the 31 codons with RSCU values greater than 1, the AUG codon encoding methionine had the highest utilization bias (*C. arborescens* RSCU: 2.99 (Fig. 4A), *C. opulens* RSCU: 2.98 (Fig. 4B)). Tryptophane had no codon usage bias among these 31 codons (only one codon). Except for UUG, which encodes leucine, and AUG, which encodes methionine, the remaining codons terminated in A (12) or U (16) (Table S4).

Sequence divergence analysis

Previous research has demonstrated that highly variable loci in the plastid genome can be used to investigate molecular markers [13]. Therefore, the software DNAsp6 [37] was used to calculate the nucleotide diversi (Pi) in order to identify highly variable regions in the chloroplast genomes of *C. arborescens* and *C. opulens*. According to the results of sliding window analysis, the Pi values of the two plants ranged from 0 to 0.05516, with an average value of approximately 0.0067655 (Fig. 5), indicating that the chloroplast genome sequences of the same genus have few distinctions and a high degree of similarity. *rpoC2-rps2*, *accD-cemA*, *rps18-clpP*, *rpoA-rpl36*, and *rpl2-rpl23* were determined to be the most probable highly variable regions based on the pi values of 111 different genes. Furthermore, the *rpoA-rpl36* region has the highest pi value, followed by the *rps18-clpP* region.

To demonstrate the distinct chloroplast genome sequence levels in *C. arborescens* and *C. opulens*. *Caragana*, including *C. arborescens*, *C. opulens*, *C. kozlowii*, *C. rosea*, *C. microphylla*, and *C. korshinskii*, had their whole plastid genome sequences compared to that of *C. jubata* (Fig. 6). Extremely low sequence divergence among species suggests that the CP genome was more conservative. IGS (*matK-rbcL*), IGS (*psbM-petN*), IGS (*atpA-psbI*), IGS (*petA-psbL*), IGS (*psbE-petL*), and IGS (*rps7-rps1 2*) contain significant differences among *Caragana* plants. Additionally, the majority of protein-coding regions were highly conserved, with the exception of a few (*accD*, *ycf2*, and *rps7*). This indicates that IGS is responsible for the accelerated evolution of *Caragana* species.

Phylogenetic development analysis

For determining the phylogenetic position of *Caragana* in the family Fabaceae, we have made multiple sequence matches using 86 protein sequences commonly found in 23 plastids. Except for the genus *Caragana*, the remaining 8 genus included *Wisteria*(1), *Glycyrrhiza*(2), *Astragalus*(1), *Calophaca*(1), *Cicer*(1), *Medicago*(3), *Trifolium*(3), and *Lathyrus*(4). Numbers in brackets indicate the number of species in the relevant group.

Based on the chloroplast genomes of 22 Fabaceae and Arabidopsis thaliana (outgroup), phylogenetic trees were constructed by Bayesian and maximum likelihood methods. The similar topology of the phylogenetic trees acquired by the two methods. Phylogenetic analysis revealed that all samples were classified into three main branches. Undoubtedly, *C. arborescens* and *C. opulens* were related and clustered together as the other four *Caragana* species, with a bootstrap value of 100%(Fig. 7). The two pairs showed a closer relationship: *C. microphylla* and *C. korshinskii*, and *C. opulens* and *C. rosea*. From the results, it is noteworthy that the genus *Astragalus* and *Caragana* were relatively closely related(bootstrap:100%) and categorized into *Subtrib*. Astragalinae. This result was consistent with previous studies[2].

Discussion

Caragana is one of the superb forages native to northwest China and certain plateau regions, and it has significant application value for the enhancement of natural pastures and the establishment of forage bases. Due to their adaptability to drought conditions, *Caragana* plants are extensively cultivated due to their resistance to drought, aridity, cold, and heat. Using the Illumina platform, we sequenced the complete CP genomes of two *Caragana* plants for this study. By assembling and annotating these genomes, more detailed information was obtained. Two plastids ranged in size from 129,473 to 132,815 (bp)base pairs and were found in *C. arborescens* and *C. opulens*, respectively. Other *Caragana* have comparable gene structures. In some species, the chloroplast genome reportedly lacks *ycf2*, *rpl23*, and *accD* [38–40], whereas these genes were present in *Caragana*. Throughout plant evolution, several genes have been lost from the plastid genome. Previous research had demonstrated that the r*pl22* and *infA* genes were lost in some or all legumes [41], that *infA* is an abnormally unstable flowering plant cp gene, and that *rpl22* is a gene encoding ribosomal protein *CL22* that was lost in cpDNA and relocated to the nucleus [42, 43]. Similarly, the *infA* and *rpl22* loci in *C. arborescens* and *C. opulens* were also deleted in this study. Recent research has demonstrated, however, that the *infA* gene was present in the chloroplast genomes of *C. jubata*, *C. erinacea*, and *C. bicolor*[2]. This phenomenon suggests that *infA* genes may not be lost or transferred to organelles in some *Caragana* plants. Due to the limited quantity of chloroplast genome data in *Caragana* plants, this conclusion must be confirmed through an extensive number of experimental studies.

As with the majority of plant species, the plastids of two *Caragana* species are conserved and no rearrangements have occurred. Multiple *Caragana* species, such as *C. microphylla*, *C. erinacea*, and *C. intermedia*, have reportedly lost their IR region [2, 11, 12]. Similarly, the CP genomes of *C. arborescens* and *C. opulens* examined in this study lacked the IR region, and the two plants share a high degree of similarity in terms of genomic structure, gene deletion, genomic size, gene types, repeat sequence distribution, etc. Moreover, cpDNA G/C content is a key determinant of inter-specific affinity [2], and the DNA G/C content of the two Caragana species discussed in this paper is extremely similar. Numerous repetitive sequences were identified in the plastid genomes of two plants. These sequences are significant genetic markers and are intimately associated with the origin and evolution of species[44]. Four types of repeats were identified in this study: complement repeats (C), reverse repeats (R), forward repeats (F), and palindromic repeats (P). The dispersed repeats were longer than 30 base pairs, and the repeat sequence length of two Caragana plants varied between 30 and 472 base pairs. In addition, SSRs are regarded as essential molecular markers for population variation research and are widely employed to assess genetic diversity, phylogenetic relationships, and evolution [45]. In total, the CP genomes of two *Caragana* plants contain between 265 and 277 SSRs with a significant A/T bias. In this study, it was determined that the majority of SSR types are single nucleotide repeats and that non-coding regions (IGSs) contain the most SSRs. Similar circumstances have been observed in other *Caragana* species, including *C. rosea*, *C. microphylla*, *C. korshinskii*, and *C. kozlowii* [27]. These repeat sequences provide a crucial starting point for the development of genetic indicators for Caragana species and can be utilized for phylogenetic and ecological research.

It is known that the codon utilization preference mirrors the species of origin and the mutational model. The study of codon bias patterns in chloroplast genomes can shed light on plant phylogenetic relationships, gene expression mechanisms, and molecular evolution [36]. Leucine (Leu) is the most abundant amino acid in *C. arborescens* and *C. opulens* (mean of 1969), and the same trend has been observed in other *Caragana* species. In addition, our research revealed that the majority of synonymous codons preferred for RSCU values terminated in A/U, resulting in a high AT content in the gene. Based on the preceding analysis, we infer that natural selection and gene mutation may be to blame. Codon preference and utilization patterns reflect the evolutionary relationship between species to a limited extent [46], but additional research is required.

Meanwhile, we identified five intergenic spacer regions (IGSs) with relatively high differentiation values (pi > 0.01037): *rpoC2-rps2*, *accD-cemA*, *rps18-clpP*, *rpoA-rpl36*, and *rpl2-rpl23*. In addition, fragmentary sequences of the *ycf1*, *rps3*, and *rps7* genes exhibited comparatively high nucleotide diversity. These variable regions could potentially function as DNA barcode labels for phylogenetic relationships, species recognition, and population genetics research [47–49]. The sequence variations of six assembled *Caragana* plants were then compared. Analysis of the comparative CP genomes revealed that the code regions were more conserved than the untranslated regions, corroborating findings from other *Caragana* species.

The phylogenetic analysis of 23 Fabaceae species simultaneously revealed the phylogenetic position of *Caragana* and the relationship between *Caragana* and closely related species. The evolution of the plastid genome (nucleotide changes and structural changes) has been elucidated by advances in phylogenetic analysis [50, 51], according to previous research. *Caragana* species were monophyletic, and *C. arborescens*, *C. opulens*, *C. kozlowii*, *C. rosea*, *C. microphylla* and *C. korshinskii* were distinguishable from other species. Our findings may serve as a guide for future research into the phylogenetic evolution of *Caragana* plants and the creation of novel molecular markers. Our findings augment the chloroplast genome database of the *Caragana* genus.

Conclusions

In this study, we constructed phylogenetic relationships based on the chloroplast genome sequences of *C. arborescens*, *C. opulens*, and 23 *Legume* species belonging to IRLC. The long repeats, SSRs loci, codon usage bias and five hypervariable regions identified in our paper are helpful for future works, such as the development of new molecular markers, population genetics and phylogenetic analysis. We deeply analyzed sequences and structure information of the chloroplast genomes of two *Caragana* plants, as well as the genetic evolution position and evolutionary relationship with other genus *Caragana*, which provided a data basis for more in-depth and comprehensive study of *Caragana* species identification, genetic diversity and phylogenetic research. At the same time, the chloroplast genome database of *Caragana* plants were also effectively enriched.

Materials and methods

DNA Extraction, library construction, and sequencing

C. arborescens and *C. opulens* leaves were collected in Qinghai Province (China) at the following coordinates: *C. arborescens*: 36°43'24.80"N, 101°44'54.11"E; *C. opulens*: 37°36'53.34"N, 10°19'18.63"E. The improved cetyltrimethylammonium bromide (CTAB) method was used to extract whole genome DNA from fresh leaves of *Caragana* plants [52]. The sample's genomic DNA needs to be evaluated. When the test was successful, the mechanical damage method was utilized to ultrasonically fragment the sample DNA, purify the genomic DNA, and stop the repair process. The DNA fragment size was subsequently determined using agarose gel electrophoresis, and the sequencing library was generated using PCR amplification. The qualified library was sequenced utilizing the Illumina Novaseq platform, with 150 bp pair-end reads.

Gene annotation and sequence analyses

This study utilized Trimmomatic [53]to remove low-quality data from the original data. The chloroplast genome sequence was then assembled using SPAdes [54]software to obtain its seed sequence, Kmer analysis of the seed sequences to obtain the contigs, and SSPACE v2 [55] software to connect the contigs and obtain the scaffold sequences. Gapfiller v2.1.1 [56] software was used to supplement the GAP found in the scaffold sequence to assure the integrity of the pseudogenome sequence. After adjusting the corrected pseudogenome sequence, complete CP genome sequences of *C. arborescens* and *C. opulens* were ultimately obtained. Blast(https://blast.ncbi.nlm.nih.gov/Blast.cgi) was used to derive the CP genome annotation results of the two plants. Hmmer (http://www.hmmer.org) and ARAGORN [57] (http://ogdraw.mpimp-golm.mpg.de/index.shtml) were used to obtain the annotation information for rRNA and tRNA, respectively. OGDRAW [58] (http://ogdraw.mpimp-gol.m.mpg.de/index.html)was used to plot the CP genome maps of *C. arborescens* and *C. opulens*. MT211962, OQ656872 are the NCBI accession numbers for the whole chloroplast genomic data reported in this investigation.

Repeat structure, SSRs and codon usage analysis

Vmatch Web [59] identified repeats (forward repeats, palindromic repeats, reverse repeats, and complementary repeats). MISA [60] software was used to identify SSRs in two *Caragana* plants with the following search parameters: mononucleotides set to ≥ 10 repeat units, dinucleotides ≥ 8 repeat units, trinucleotides, tetranucleotides, pentanucleotides and hexanucleotides ≥ 3 repeat units. CodonW1.4.2 calculated the relative synonymous codon usage (RSCU) values of protein-coding genes using the default settings.

Comparative genome analysis

C. arborescens and *C. opulens* whole cp genome sequences were compared with those of *C. kozlowii*, *C. rosea*, *C. microphylla*, and *C. korshinskii* using mVISTA (shuffle-lagan mode). As a reference, *C. jubata* plastid was labeled. 111 gene sequences between *C. arborescens* and *C. opulens* were aligned using MEGA7 [61]. To calculate the nucleotide diversity (Pi) values with the software DnaSP6, the following parameter configurations are utilized: Normal parameters, 200-bp step size and 600-bp window length, appeared in the document [62].

Phylogenetic analysis

Phylogenic tree was established using plastome sequences of 20 species (pertaining to IRLC)downloaded from NCBI database and two species sequenced in this study, with *Arabidopsis* as an outgroup. All 23 complete chloroplast genomes were aligned using the tool MAFFT(default parameters), and the aligned sequences were optimized using MACSE. Bayesian inference (BI) and maximum likelihood (ML) methods were used to construct phylogenetic trees. MrBayes was used for BI analysis, with the model selected using Modelfinder [63] and the nucleotide substitution model set to GTR + F + I + G4. The MrBayes analysis was set to run for 1,000 cycles, and the first 25% of the cycles were removed as burn-in. The average standard deviation of splitfrequencies was set to > 0.01 [64]. IQ-TREE was used for ML analysis, with the automatic partitioning module and bootstrap analysis set to 1,000 repetitions to evaluate the confidence of the branches.

Declarations

Acknowledgments

This research was funded by Innovation Platform for the Development and Construction of Special Project of Qinghai Province (Grant Number 2021-ZJ-T05) and National Natural Science Foundation (No. 32160386) of China.

Authors' contributions

LiE Liu was the study's experimental designer and executor, completing data analysis and writing the first draft of the paper. HongYan Li, JiaXin Li, and XinJuan Li all contributed to the experimental design and analysis of the results. Na Hu and Honglun Wang assisted in sample collection and species identification. Wu Zhou was the project developer and leader, guiding the experimental design, data analysis, and paper writing and revision. The final text has been read and approved by all authors.

Funding

Innovation Platform for the Development and Construction of Special Project of Qinghai Province (Grant Number 2021-ZJ-T05) and National Natural Science Foundation (No. 32160386) of China.

Availability of data and materials

The original sequencing data have been submitted to the NCBI database and received GenBank accession numbers MT211962(C. arborescens), OQ656872(C. opulens). The data used in this study are already entirely in the public domain (https://www.ncbi.nl m .nih.gov).

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

Address: [1] School of Ecological and Environmental Engineering, Qinghai University, Xining, China, [2] Key Laboratory of Tibetan Medicine Research, Northwest Institute of Plateau Biology, Chinese Academy of Sciences, Xining 810008, China.

Email: LiE Liu - ys210854000328@qhu.edu.cn; HongYan Li - ys210713000111@qhu.edu.cn; lijiaxin@qhu.edu.cn; XinJuan Li - ys220713000116@qhu.edu.cn; Na Hu - huna@nwipb.cas.cn; Honglun Wang - hlwang@nwipb.cas.cn; Wu Zhou* - zhouwu870624@qhu.edu.cn

* Corresponding author.

References

1. Meng Q, Niu Y, Niu X, Roubin RH, Hanrahan JR. **Ethnobotany, phytochemistry and pharmacology of the genus Caragana used in traditional Chinese medicine**. *Journal of ethnopharmacology* 2009; **124**(3):350-368.
2. Yuan M, Yin X, Gao B, Gu R, Jiang G. **The chloroplasts genomic analyses of four specific Caragana species**. *PloS one* 2022; **17**(9):e0272990.
3. Kang HM, Chen K, Bai J, Wang G. **Antioxidative system's responses in the leaves of six Caragana species during drought stress and recovery**. *Acta Physiologiae Plantarum* 2012; **34**(6):2145-2154.
4. Fei, Ma, Xiaofan, Na, Tingting, Xu. **Drought responses of three closely relatedCaraganaspecies: implication for their vicarious distribution**. *Ecology and Evolution* 2016; **6**(9):2763-2773.
5. Moukoumi JL, Hynes RK, Dumonceaux TJ, Town J, Bélanger N. **Characterization and genus identification of rhizobial symbionts from Caragana arborescens in western Canada**. *Canadian Journal of Microbiology* 2013; **59**(6):399-406.
6. Sun X, Ma J, Li C, Zang Y, Tian J, Li L, Li Y, Ye F, Zhang D. **Hypoglycemic oligostilbenes from the stems of Caragana sinica**. *Bioorganic chemistry* 2023; **134**:106458.

7. Kordyum E, Bilyavska N. **Structure and biogenesis of ribonucleoprotein bodies in epidermal cells of Caragana arborescens L**. *Protoplasma* 2018; **255**(2):709-713.
8. Ma C, Gao Y, Li Q, Guo H, Zhang J, Shi Y. **Water regulation characteristics and stress resistance of Caragana opulens population in different habitats of Inner Mongolia plateau**. *Ying yong sheng tai xue bao = The journal of applied ecology* 2006; **17**(2):187-191.
9. Luo HF, Zhang LP, Hu CQ. **ChemInform Abstract: Five Novel Oligostilbenes from the Roots of Caragana sinica**. *ChemInform* 2010; **32**(37).
10. Pan L, Zhang T, Yu M, Shi M, Zou Z. **Bioactive-guided isolation and identification of oligostilbenes as anti-rheumatoid arthritis constituents from the roots of Caragana stenophylla**. *Journal of Ethnopharmacology* 2021;114134.
11. Liu BB, Duan N, Zhang HL, Liu S, Shi JW, Chai BF. **Characterization of the whole chloroplast genome of Caragana microphylla Lam (Fabaceae)**. *Conservation Genetics Resources* 2016; **8**(4):371-373.
12. Zhang ZL, Ma LY, Yao HB, Yang X, Luo JH, Gong X, Wei SY, Li QF, Wang W, Sun HB. **Complete chloroplast genome of Caragana intermedia (Fabaceae), an endangered shrub endemic to china**. *Conservation Genetics Resources* 2016; **8**(4):1-3.
13. Mei J, Haimei C, Shuaibing H, Liqiang W, Amanda C, Chang L. **Sequencing, Characterization, and Comparative Analyses of the Plastome of Caragana rosea var. rosea**. *International Journal of Molecular Sciences* 2018; **19**(5):1419.
14. Kim K, Lee H. **Complete chloroplast genome sequences from Korean ginseng (Panax schinseng Nees) and comparative analysis of sequence evolution among 17 vascular plants**. *DNA research : an international journal for rapid publication of reports on genes and genomes* 2004; **11**(4):247-261.
15. Peirong L, Shujiang Z, Fei L, Shifan Z, Hui Z, Xiaowu W, Rifei S, Guusje B, Borm TJA. **A Phylogenetic Analysis of Chloroplast Genomes Elucidates the Relationships of the Six Economically Important Brassica Species Comprising the Triangle of U**. *Frontiers in Plant Science* 2017; **8**:111-.
16. Shen X, Wu M, Liao B, Liu Z, Bai R, Xiao S, Li X, Zhang B, Xu J, Chen S. **Complete Chloroplast Genome Sequence and Phylogenetic Analysis of the Medicinal Plant Artemisia annua**. *Molecules (Basel, Switzerland)* 2017; **22**(8).
17. Neuhaus, HE, Emes, MJ. **NONPHOTOSYNTHETIC METABOLISM IN PLASTIDS**. *Annual Review of Plant Physiology & Plant Molecular Biology* 2000; **51**(1):111-111.
18. Allen JF. **Why chloroplasts and mitochondria contain genomes**. *Hindawi Publishing Corporation* 2003; (1).
19. Zhang T, Xing Y, Xu L, Bao G, Kang T. **Comparative analysis of the complete chloroplast genome sequences of six species of Pulsatilla Miller, Ranunculaceae**. *Chinese Medicine* 2019; **14**(1).
20. Somaratne Y, Guan DL, Wang WQ, Zhao L, Xu SQ. **The Complete Chloroplast Genomes of Two Lespedeza Species: Insights into Codon Usage Bias, RNA Editing Sites, and Phylogenetic Relationships in Desmodieae (Fabaceae: Papilionoideae)**. *Plants* 2020; **9**(1).
21. Nai-hu HYLC-IMCW. **Chloroplast DNA and Its Application to Plant Systematic Studies**. *Chinese Bulletin of Botany* 1994; **11**(02):11-25.
22. Yun S, Yan C, Lv J, Jin X, Zhu S, Li MF, Chen N. **Development of Chloroplast Genomic Resources for Oryza Species Discrimination**. *Frontiers in Plant ence* 2017; **8**:1854.

23. Jiao Y, Ming Y, Chuan N, Xiong-Feng M, Zhong-Hu L. **Comparative Analysis of the Complete Chloroplast Genome of Four Endangered Herbals of Notopterygium**. *Genes* 2017; **8**(4):124.
24. Palmer JD, Thompson WF. **Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost**. *Cell* 1982; **29**(2):537-550.
25. Sabir J, Schwarz E, Ellison N, Zhang J, Baeshen NA, Mutwakil M, Jansen R, Ruhlman T. **Evolutionary and biotechnology implications of plastid genome variation in the inverted-repeat-lacking clade of legumes**. *Plant Biotechnology Journal* 2014; **12**(6):743-754.
26. Lei W, Ni D, Wang Y, Shao J, Liu C. **Intraspecific and heteroplasmic variations, gene losses and inversions in the chloroplast genome of Astragalus membranaceus**. *Scientific Reports* 2016; **6**:21669.
27. Moghaddam M, Ohta A, Shimizu M, Terauchi R, Kazempour-Osaloo S. **The complete chloroplast genome of Onobrychis gaubae (Fabaceae-Papilionoideae): comparative analysis with related IR-lacking clade species**. *BMC Plant Biology* 2022.
28. Zhu S, Liu A, Xie X, Xia M, Chen H. **Wisteriopsis reticulataCharacterization of the complete chloroplast genome of (Fabaceae): an IRLC legumes**. *Mitochondrial DNA Part B, Resources* 2022; **7**(6):1137-1139.
29. Zhumanova K, Lee G, Baiseitova A, Shah AB, Park KH. **Inhibitory mechanism of O-methylated quercetins, highly potent β-secretase inhibitors isolated from Caragana balchaschensis (Kom.) Pojark**. *Journal of Ethnopharmacology* 2021; **272**(421):113935.
30. Raman G, Park KT, Kim JH, Park SJ. **Characteristics of the completed chloroplast genome sequence of Xanthium spinosum: Comparative analyses, identification of mutational hotspots and phylogenetic implications**. *BMC Genomics* 2020; **21**(1).
31. Jianguo Z, Yingxian C, Xinlian C, Ying L, Zhichao X, Baozhong D, Yonghua L, Jingyuan S, Hui Y. **Complete Chloroplast Genomes of Papaver rhoeas and Papaver orientale: Molecular Structures, Comparative Analysis, and Phylogenetic Analysis**. *Molecules* 2018; **23**(2):437.
32. Jo YD, Park J, Kim J, Song W, Hur CG, Lee YH, Kang BC. **Complete sequencing and comparative analyses of the pepper (Capsicum annuum L.) plastome revealed high frequency of tandem repeats and large insertion/deletions on pepper plastome**. *Plant Cell Reports* 2011; **30**(2):217-229.
33. Sloan D, Triant D, Forrester N, Bergner L, Wu M, Taylor D. **A recurring syndrome of accelerated plastid genome evolution in the angiosperm tribe Sileneae (Caryophyllaceae)**. *Molecular phylogenetics and evolution* 2014; **72**:82-89.
34. Ebert D, Peakall R. **Chloroplast simple sequence repeats (cpSSRs): technical resources and recommendations for expanding cpSSR discovery and applications to a wide array of plant species**. *Molecular Ecology Resources* 2010; **9**.
35. Kuang DY, Wu H, Wang YL, Gao LM, Lu L. **Complete chloroplast genome sequence of Magnolia kwangsiensis (Magnoliaceae): implication for DNA barcoding and population genetics**. *Genome* 2011; **54**(8):663-673.
36. Parvathy S, Udayasuriyan V, Bhadana V. **Codon usage bias**. *Molecular biology reports* 2022; **49**(1):539-565.
37. Rozas J, Ferrer-Mata A, Sánchez-DelBarrio J, Guirao-Rico S, Librado P, Ramos-Onsins S, Sánchez-Gracia A. **DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets**. *Molecular biology and evolution*

2017; **34**(12):3299-3302.

38. Jansen RK, Cai Z, Raubeson LA, Daniell H, Boore JL. **Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns**. *Proceedings of the National Academy of Sciences* 2008; **104**(49):19369-19374.
39. Oliver M, Murdock A, Mishler B, Kuehl J, Boore J, Mandoli D, Everett K, Wolf P, Duffy A, Karol K. **Chloroplast genome sequence of the moss Tortula ruralis: gene content, polymorphism, and structural arrangement relative to other green plant chloroplast genomes**. *BMC genomics* 2010; **11**:143.
40. 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 Molecular Biology* 2011; **76**(3-5):273-297.
41. Yen L, Kousar M, Park J. **Desmodium stryacifoliumComparative Analysis of Chloroplast Genome of with Closely Related Legume Genome from the Phaseoloid Clade**. *International journal of molecular sciences* 2023; **24**(7).
42. Gantt JS, Baldauf SL, Calie PJ, Weeden NF, Palmer JD. **Transfer of rpl22 to the nucleus greatly preceded its loss from the chloroplast and involved the gain of an intron**. *Embo Journal* 1991; **10**(10):3073-3078.
43. Millen RS, Olmstead RG, Adams KL, Palmer JD, Lao NT, Heggie L, Kavanagh TA, Hibberd JM, Gray JC, Morden CW. **Many Parallel Losses of infA from Chloroplast DNA during Angiosperm Evolution with Multiple Independent Transfers to the Nucleus**. *Plant Cell* 2001; **13**(3):645-658.
44. Xie D, Yu Y, Deng Y, Li J, Liu H, Zhou S, He X. **UrophysaComparative Analysis of the Chloroplast Genomes of the Chinese Endemic Genus and Their Contribution to Chloroplast Phylogeny and Adaptive Evolution**. *International journal of molecular sciences* 2018; **19**(7).
45. Varshney RK, Sigmund R, Borner A, Korzun V, Stein N, Sorrells ME, Langridge P, Graner A. **Interspecific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice**. *Plant Science* 2005; **168**(1):195-202.
46. Chen M, Zhang M, Liang Z, He Q. **UncariaCharacterization and Comparative Analysis of Chloroplast Genomes in Five Species Endemic to China**. *International journal of molecular sciences* 2022; **23**(19).
47. Henriquez CL, Abdullah, Ahmed I, Carlsen MM, Mckain MR. **Evolutionary dynamics of chloroplast genomes in subfamily Aroideae (Araceae)**. *Genomics* 2020; **112**(3):2349-2360.
48. Zheng G, Wei L, Ma L, Wu Z, Gu C, Chen K. **Comparative analyses of chloroplast genomes from 13 Lagerstroemia (Lythraceae) species: identification of highly divergent regions and inference of phylogenetic relationships**. *Plant molecular biology* 2020; **102**(6):659-676.
49. Park I, Song J, Yang S, Choi G, Moon B. **A Comprehensive Study of the Genus Sanguisorba (Rosaceae) Based on the Floral Micromorphology, Palynology, and Plastome Analysis**. *Genes* 2021; **12**(11):1764-.
50. Saski C, Lee SB, Daniell H, Wood TC, Tomkins J, Kim HG, Jansen RK. **Complete Chloroplast Genome Sequence of Glycine max and Comparative Analyses with other Legume Genomes**. *Plant Molecular Biology* 2005; **59**(2):309-322.
51. Haberle RC, Fourcade HM, Boore JL, Jansen RK. **Extensive Rearrangements in the Chloroplast Genome of Trachelium caeruleum Are Associated with Repeats and tRNA Genes**. *Journal of Molecular Evolution* 2008; **66**(4):350-361.

52. Dierckxsens N, Mardulyn P, Smits G. **NOVOPlasty: De novo assembly of organelle genomes from whole genome data**. 2016.

53. Bolger AM, Marc L, Bjoern U. **Trimmomatic: a flexible trimmer for Illumina sequence data**. *Bioinformatics* 2014; (15):2114-2120.

54. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD. **SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing**. *Journal of Computational Biology* 2012; **19**(5):455-477.

55. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. **Scaffolding pre-assembled contigs using SSPACE**. *Bioinformatics* 2011; **27**(4):578-579.

56. Boetzer M, Pirovano W. **Toward almost closed genomes with GapFiller**. *Genome Biology,13,6(2012-06-25)* 2012; **13**(6):R56.

57. Laslett D, Canback B. **ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences**. *Oxford University Press* 2004 (1).

58. Stephan G, Pascal L, Ralph B. **OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes**. *Nuclc Acids Research* 2019(W1);W59-W64.

59. Kurtz S. **The Vmatch large scale sequence analysis software-A Manual**. *Center for Bioinformatics* 2010; **170**(24):391–392.

60. Sebastian, Beier, Thomas, Thiel, Münch, Uwe, Scholz, Martin, Mascher. **MISA-web: a web server for microsatellite prediction**. *Bioinformatics (Oxford, England)* 2017.

61. Kumar S, Stecher G, Tamura K. **MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets**. *Molecular Biology and Evolution* 2015; **33**:1870-1874.

62. Xu F, He L, Gao S, Su Y, Xu L. **Comparative Analysis of two Sugarcane Ancestors Saccharum officinarum and S. spontaneum based on Complete Chloroplast Genome Sequences and Photosynthetic Ability in Cold Stress**. *International Journal of Molecular Sciences* 2019; **20**(15):3828-.

63. Kalyaanamoorthy S, Minh BQ, Wong T, Haeseler AV, Jermiin LS. **ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates**. *Nature Methods* 2017; **14**(6).

64. Hu G, Wang Y, Wang Y, Zheng S, Dong N. **New Insight into the Phylogeny and Taxonomy of Cultivated and Related Species of Crataegus in China, Based on Complete Chloroplast Genome Sequencing**. *Horticulturae* 2021; **7**(9):301.

Figures

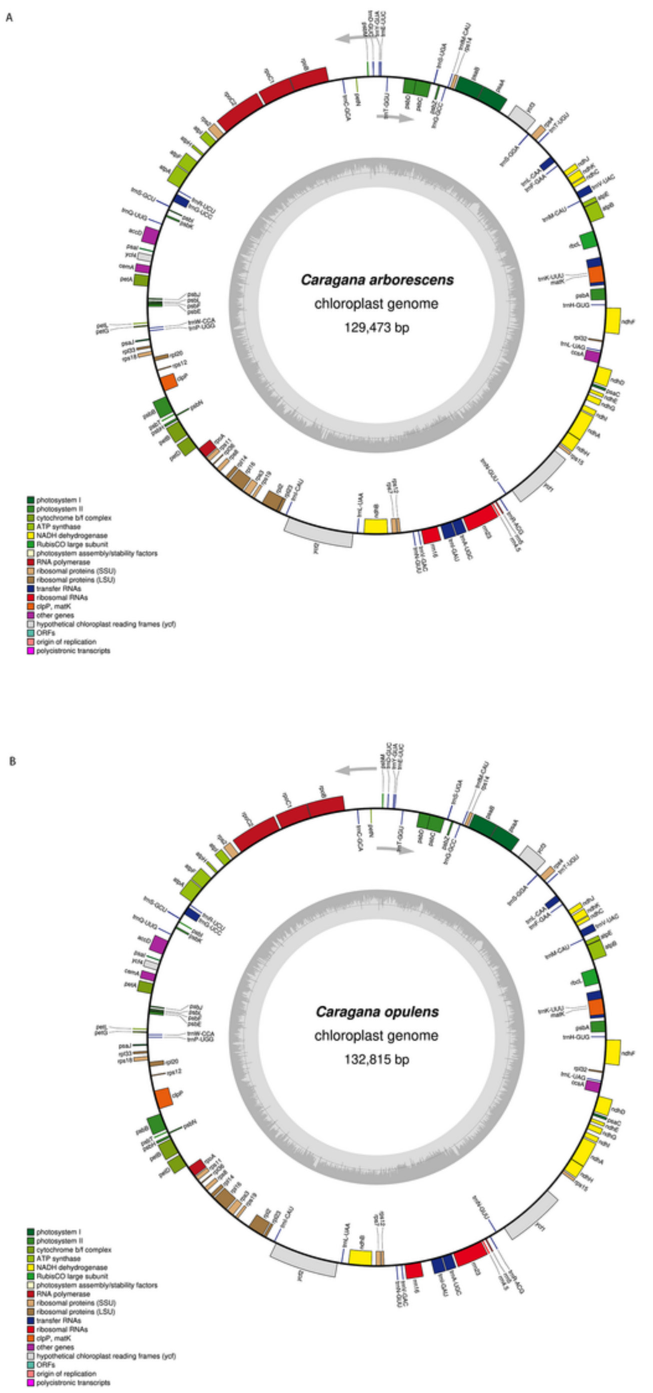

Figure 1

The diagram illustrates the chloroplast gene maps of *C. arborescens* **A** and *C. opulens* **B**. The genes within the circle are transcribed in a clockwise direction, whereas the genes outside the circle are transcribed in the opposite direction. Different color codes are used to depict functionally distinct genomes. In addition, the change in GC content in the inner ring is shown in light gray, whereas the change in AT content is shown in dark gray.

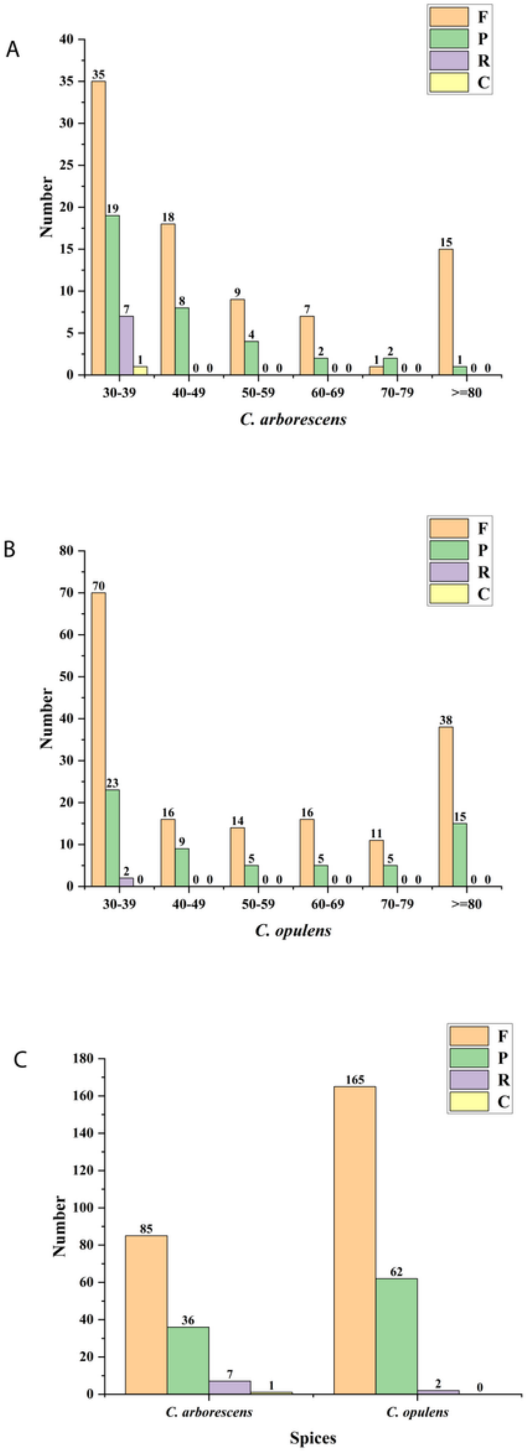

Figure 2

Repetitive sequences in the chloroplast genomes of *C. arborescens* and *C. opulens*.

A Numbers of four types of repeats found in *C. arborescens*; **B** Numbers of four types of repeats found in *C. opulens*; **C** The total number of four types of repeat sequences in *C. arborescens* and *C. opulens*.

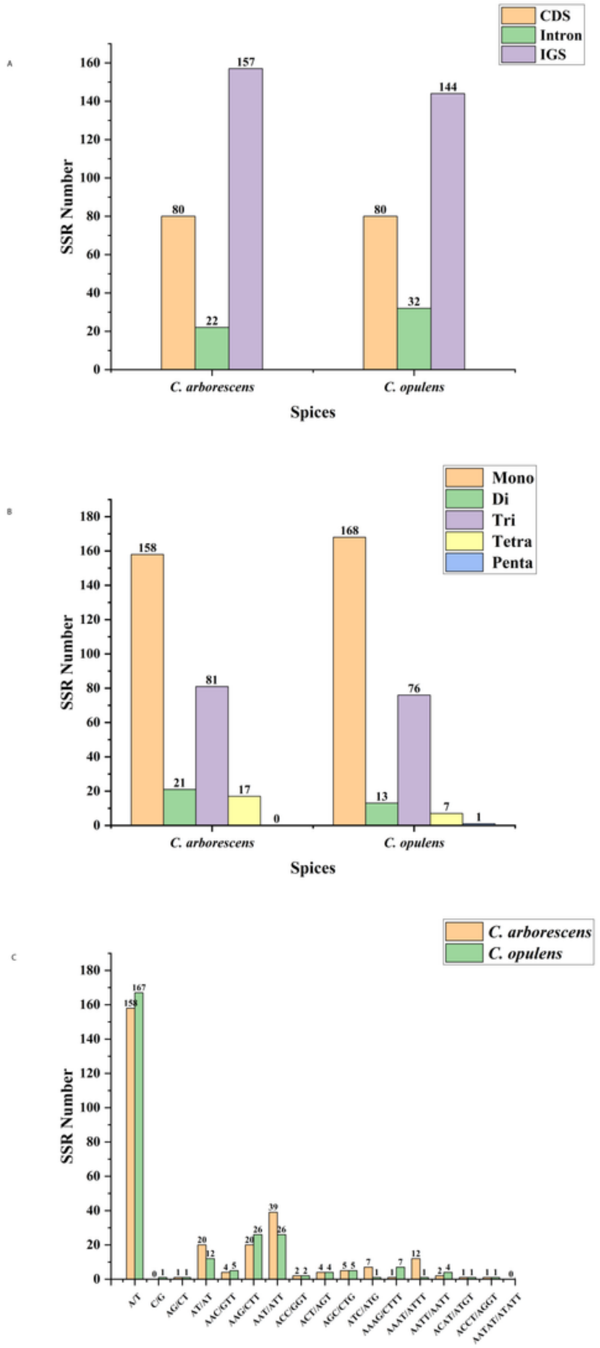

Figure 3

Simple sequence repeats (SSRs) identified in the plastid genomes of two *Caragana* species.

A The number of SSRs was found in coding (CDS), and intronic regions, intergenic (IGS), Respectively; **B** the amount of different SSR types found in the two genomes; **C** number of SSRs determined in different repetition types.

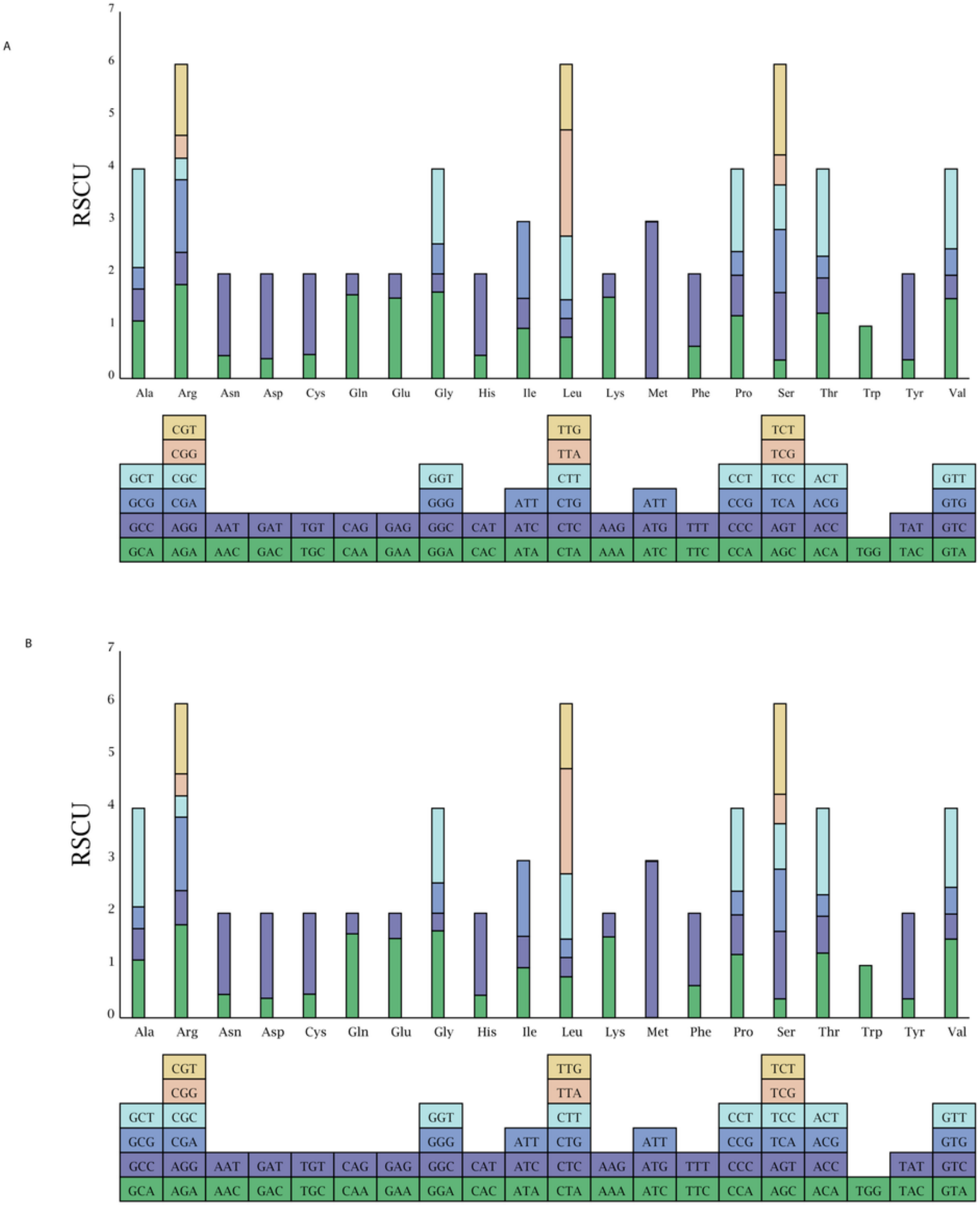

Figure 4

Amino acid frequencies of the chloroplast genomes of *C. arborescens* **A** and *C. opulens* **B**. The squares below represent all the codons that encode each type of amino acid; the height of the column above represents the total sum of RSCU values for all codons; the height of each column represents the RSCU value for each codon.

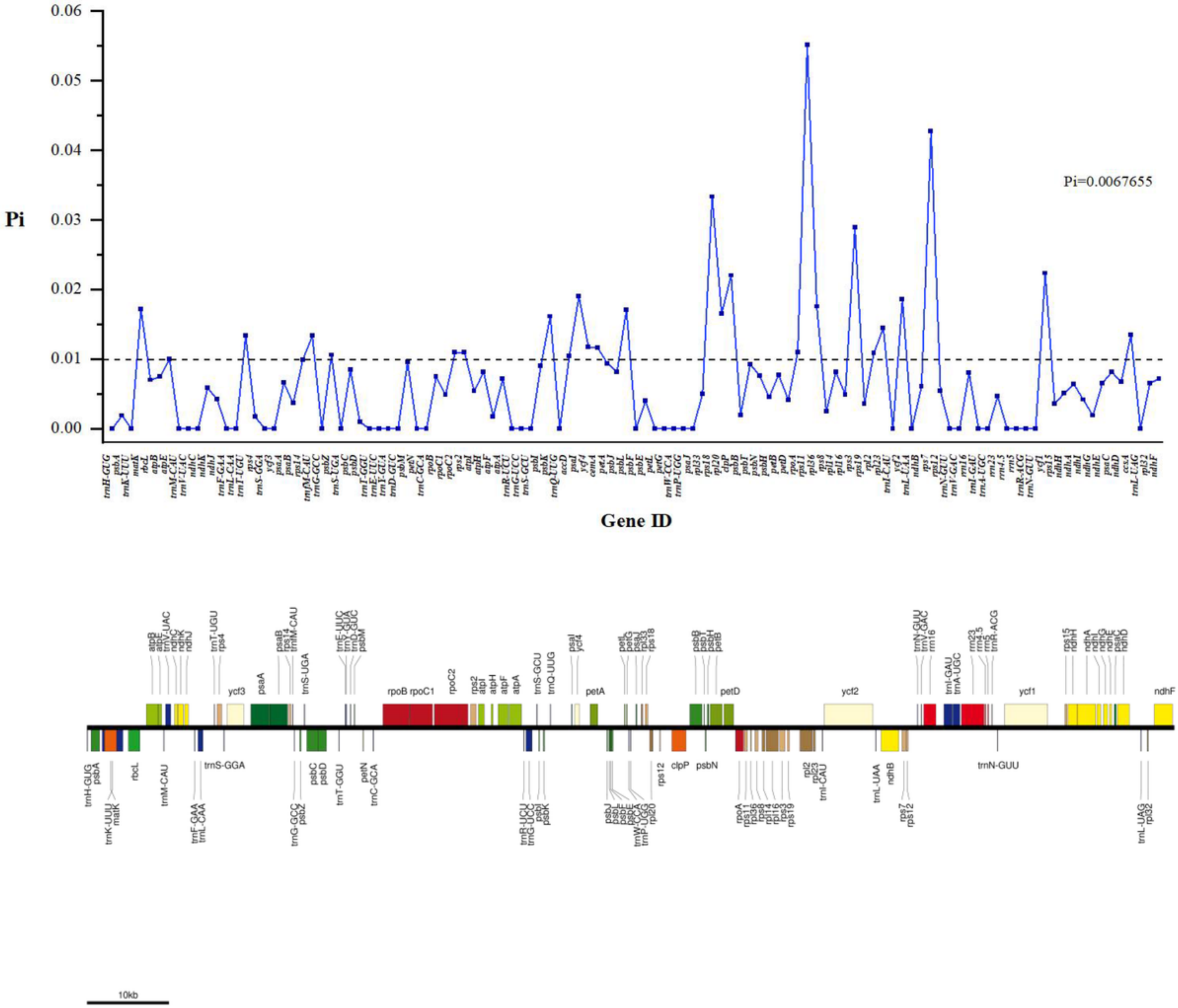

Figure 5

Nucleotide variability (Pi) values among 111 genes of two *Caragana* plants. The linear gene graph spectrum of *Caragana*species is given below.

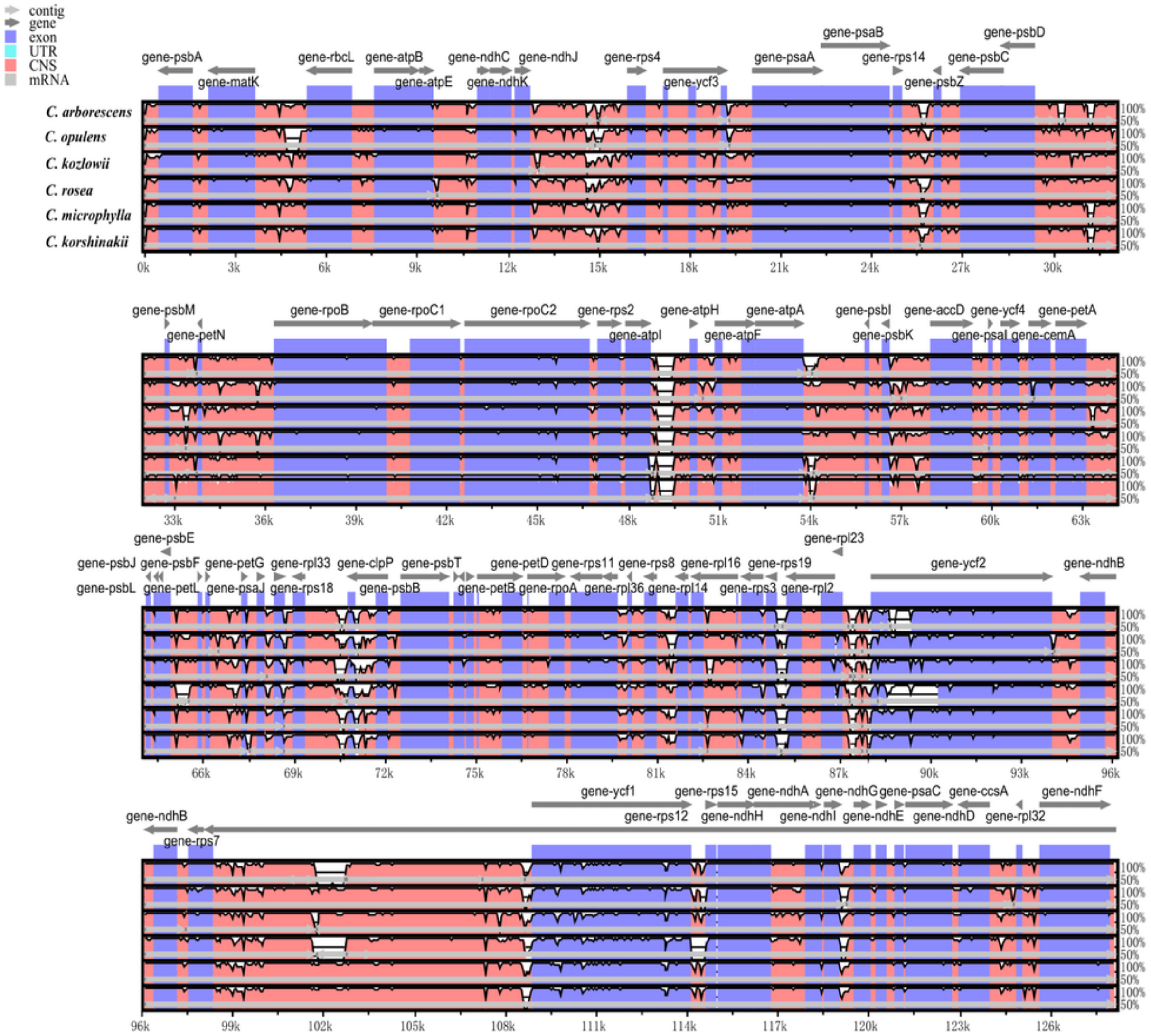

Figure 6

The chloroplast genome differences of six *Caragana*species were compared by mVISTA. The gray arrow in the figure indicates the direction of gene translation; The x-axis represents the coordinates in the CP genome; The y-axis represents the percentage between 50% and 100%; Blue indicates protein coding (exon); Light green indicates untranslated region (UTR); Orange indicates conserved non-coding sequences (CNSs).

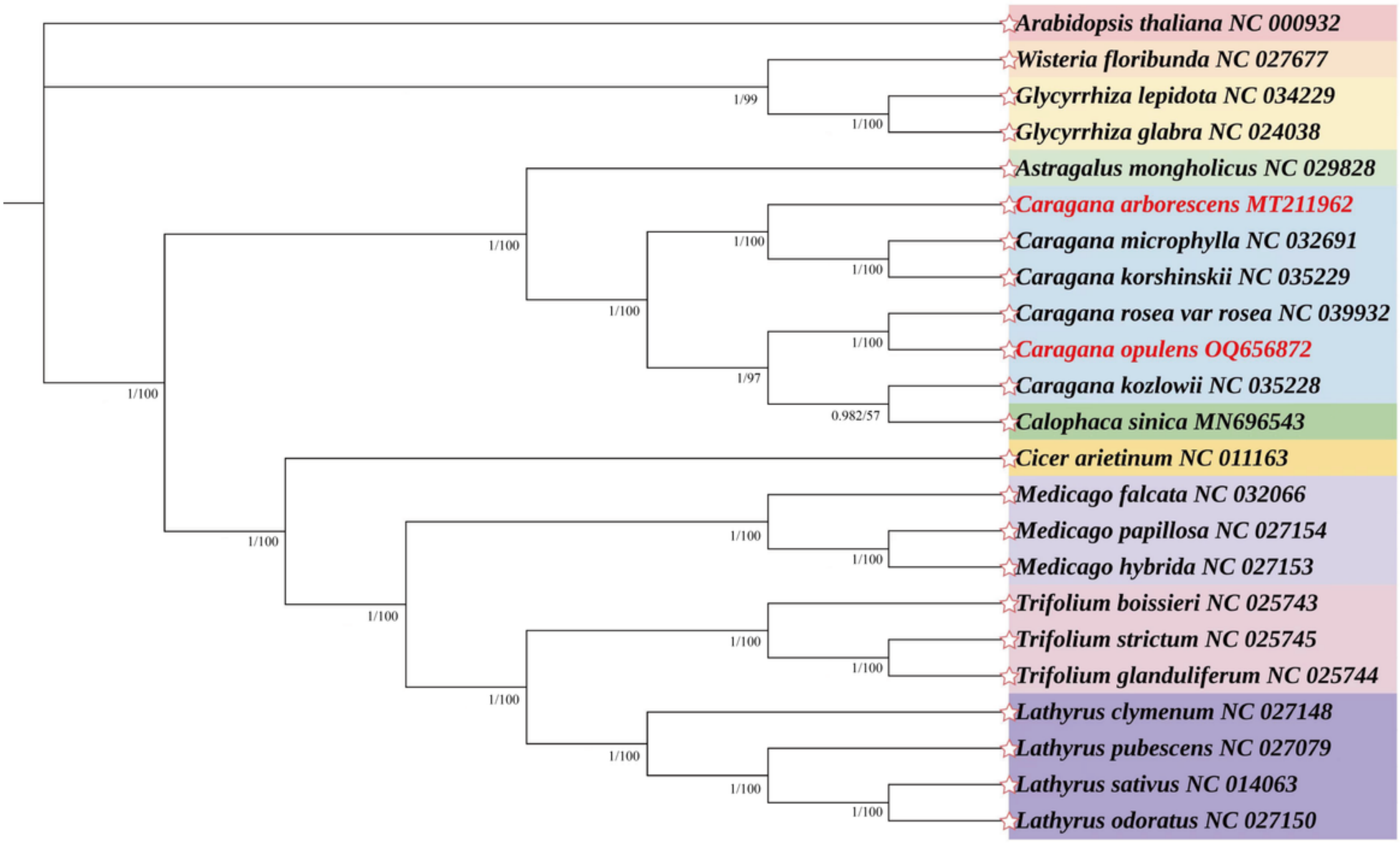
Arabidopsis thaliana NC 000932
Wisteria floribunda NC 027677
1/99
Glycyrrhiza lepidota NC 034229
1/100
Glycyrrhiza glabra NC 024038
Astragalus mongholicus NC 029828
Caragana arborescens MT211962
1/100
1/100
Caragana microphylla NC 032691
1/100
Caragana korshinskii NC 035229
1/100
Caragana rosea var rosea NC 039932
1/100
Caragana opulens OQ656872
1/97
Caragana kozlowii NC 035228
1/100
0.982/57
Calophaca sinica MN696543
Cicer arietinum NC 011163
Medicago falcata NC 032066
1/100
Medicago papillosa NC 027154
1/100
1/100
Medicago hybrida NC 027153
Trifolium boissieri NC 025743
1/100
1/100
Trifolium strictum NC 025745
1/100
1/100
Trifolium glanduliferum NC 025744
Lathyrus clymenum NC 027148
1/100
Lathyrus pubescens NC 027079
1/100
Lathyrus sativus NC 014063
1/100
1/100
Lathyrus odoratus NC 027150

Figure 7

The phylogenetic tree based on 23 plant chloroplast genomes was constructed using BI and ML. The number after the node represents the bootstrap value. The GenBank accession numbers were shown after each species. *C. arborescens* and *C. opulens*are highlighted in red.

Supplementary Files

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

- Appendixtable.docx