BMC Plant Biology

RESEARCH

Open Access

Analysis of synonymous codon usage bias in the chloroplast genome of five *Caragana*

XinJuan Li[1], LiE Liu[1], QianDan Ren[1], Tian Zhang[1], Na Hu[2], Jing Sun[2] and Wu Zhou[1*]

Abstract

Background The genus *Caragana*, known for its adaptability and high forage value, is commonly planted to rehabilitate barren land and prevent desertification. Several *Caragana* species are also used for medicinal purposes. Analysis of synonymous codon usage bias and their primary influencing factors in chloroplast genomes aims to provide insights into molecular research and germplasm innovation for *Caragana* plants.

Results The GC content of the five *Caragana* species ranged from 36.00% to 37.10%, showing a preference for codons ending in A/U, although the codon bias was weak. The screening identified nine to twelve optimal codons, but their frequency of use was low. Correlation analysis, neutrality plots, ENC plots and PR2 plots of the parameters identified two potential groups among the five species: *Caragana arborescens* and *Caragana jubata*, and *Caragana turkestanica*, *Caragana opulens* and *Caragana tibetica*. These groups showed a high level of intragroup similarity in the parameter analyses. In the RSCU cluster tree analysis, *Caragana turkestanica* and *Caragana arborescens* grouped together, while *Caragana tibetica*, *Caragana jubata* and *Caragana opulens* formed a separate clade in the CDS sequence and complete sequence phylogenetic tree analysis.

Conclusions The codon usage bias in the chloroplast genomes of the five *Caragana* species showed high similarity, suggesting that natural selection has a greater influence on codon bias than mutation. Furthermore, the identified optimal codons provide valuable insights for germplasm improvement of *Caragana* plants.

Keywords *Caragana*, Chloroplast genome, Codon usage bias, Optimal codons

Background

Caragana belongs to the subfamily Papilionoideae within the family Leguminosae and represents significant branches of this subfamily. This subfamily is characterized by its members' ability to form root nodules that fix atmospheric nitrogen, contributing significantly to soil fertility and making them well-suited for reclamation of degraded lands and combating desertification [1]. Species within the genus *Caragana* primarily appear in the form of shrubs or small trees. The genus is mainly distributed in arid and semi-arid regions of Asia and Europe, with over a hundred species. Among these, China has 66 species, of which 32 are endemic. These species are renowned for their adaptability to harsh environmental conditions, including drought, infertile soils, cold, and heat [2].

This study investigates the codon bias in the chloroplast genomes of five *Caragana* species: *Caragana arborescens*, *Caragana jubata*, *Caragana opulens*, *Caragana tibetica* and *Caragana turkestanica*. *Caragana arborescens,* which is mainly distributed in the northeast, north and northwest of China, can grow up to 4–5 m tall. It produces yellow flowers in May with seeds ripening in

*Correspondence:
Wu Zhou
zhouwu870624@qhu.edu.cn
[1] College of Eco-Environmental Engineering, Qinghai University, Xining 810016, China
[2] Qinghai Key Laboratory of Qinghai-Tibet Plateau Biological Resources, Northwest Institute of Plateau Biology, Chinese Academy of Sciences, Xining 810008, China

mid-summer and is often used as a landscape and ornamental plant [3, 4]. *Caragana jubata* is an evergreen, thorny shrub that can grow upright or prostrate and reaches a height of 0.3 to 2.0 m. It bears pink or almost white flowers and thrives in alpine bush landscapes in the north and southwest of China [5]. *Caragana opulens*, a shrub with yellow flowers, inhabits hills up to 3,400 m above sea level in the north, northwest and southwest of China [6]. *Caragana tibetica*, also called Zuomoxing, is native to western Inner Mongolia, northern Shaanxi, Ningxia, Gansu, Qinghai, western Sichuan and Tibet [7]. It grows on dry slopes and sandy areas and shows strong adaptability to sandy environments [8]. Its dense, ellipsoidal branches can collect wind-blown sand, reducing soil erosion [9]. In addition, it is an important Tibetan medicine used to promote blood circulation and relieve blood congestion. Finally, *Caragana turkestanica*, native to the Jimunai and Habahe of Xinjiang, thrives in dry scrublands and sunny slopes [10].

The species in the *Caragana* genus not only possess notable medicinal properties, being used in traditional medicine to treat various ailments such as fever, inflammation, wounds, rheumatoid arthritis, and hypertension [2], but also play a significant ecological role. Their dual role in ecological restoration and traditional medicine underscores the importance of understanding the genetic and evolutionary aspects of the genus. This diversity in habitat ensures that our findings are applicable across a broad range of environmental conditions. Additionally, the fresh plant materials can be obtained at the field research sites in Qinghai Province, ensuring the quality and quantity of the samples required for accurate sequencing and analysis. The inclusion of these five species allows for a comprehensive comparative analysis within the genus. By focusing on a manageable number of species, we can provide detailed insights into the codon usage bias and its influencing factors, which can then be extrapolated to other *Caragana* species [11]. This approach also facilitates a more robust phylogenetic analysis, highlighting the evolutionary relationships within the genus.

The genome structures of chloroplasts are highly conserved throughout evolution. Studying codon bias in chloroplast genomes improves our understanding of gene expression and regulatory mechanisms and reveals their functional roles and evolutionary dynamics. Codon bias analysis can reveal selection pressures and evolutionary trends in chloroplast genomes and provide important insights for evolutionary biology and genetic engineering [12].

Codons, which are sequences of three adjacent nucleotides, are crucial for encoding and transmitting genetic information in organisms [13]. Proteins in living organisms are composed of 20 naturally occurring amino acids, each of which is specified by at least one codon [14]. Of the 61 codons, three are stop codons, while the remaining 58 encode the 20 amino acids. Specifically, arginine, leucine, and serine each have six associated codons, while methionine and tryptophan are encoded by a single codon. The other 18 amino acids are encoded by two or more codons. Codons that encode the same amino acid are called synonymous codons [15]. Synonym codons that vary only in the third position can reduce the effects of deleterious mutations. Differences in the frequency of use of synonymous codons lead to codon bias [16], with preferentially selected codons being referred to as optimal codons [17]. Different species have different codon preferences due to factors such as natural selection, mutation pressure, gene function and gene length [18].

Wicke, S. et al. conducted extensive analyses of the chloroplast genomes in leguminous plants and found that multiple genera, including *Astragalus*, *Medicago, Pisum*, and *Vicia*, lack the inverted repeat (IR) region, a characteristic feature of the IRLC (Inverted Repeat Lacking Clade). This lack of IR regions is likely associated with genomic rearrangements and functional adaptations during their evolutionary process [19]. This studies collectively indicate that the codon usage bias in IRLC plants is not only a direct consequence of genomic structural changes but also a manifestation of functional optimization over their long evolutionary history.

Our research group sequenced, assembled and annotated the chloroplast genomes of five *Caragana* species. We then established criteria for identifying protein coding sequences (CDS) for codon use bias analysis. In the previous study on the chloroplast genome structure of *Caragana* species [20]. The chloroplast genomes of *Caragana arborescens*, *Caragana jubata, Caragana opulens*, *Caragana tibetica*, and *Caragana turkestanica* were sequenced, assembled, and annotated, revealing high conservation in gene content and order. Specifically, the initial examination of the structure of five species of *Caragana* revealed that the complete chloroplast genome length ranges from 132,815 bp in *Caragana opulens* to 128,132 bp in *Caragana jubata*. The number of genes in the chloroplast genomes is 110 for *Caragana tibetica* and *Caragana jubata*, and 111 for *Caragana arborescens*, *Caragana opulens*, *Caragana turkestanica*. All five *Caragana* include 76 protein-coding genes and 4 rRNA genes. *Caragana tibetica* and *Caragana jubata* have 30 tRNA genes, while *Caragana arborescens*, *Caragana opulens* and *Caragana turkestanica* contain 31 tRNA genes. The GC content varies between 34.30% and 34.71% across the five species. All five species belong to the Inverted Repeat Lacking Clade (IRLC), characterized

by the absence of the inverted repeat (IR) region, resulting in a more compact genome structure compared to typical angiosperm chloroplast genomes. This structural feature is consistent across all species, indicating a high degree of conservation and stability in the chloroplast genome structure [21].

Through bioinformatics methods, we analyzed the synonymous codon usage bias, the influence of natural selection and mutation, and the phylogenetic relationships of five *Caragana* species. Additionally, we identified optimal codons for germplasm improvement. The aim is to provide a comprehensive understanding of the chloroplast genome structure and codon usage bias in five *Caragana* species. By achieving these objectives, we hope to contribute to the broader knowledge of molecular mechanisms underlying environmental adaptation in *Caragana* plants and provide a foundation for future genetic engineering and conservation efforts.

Methods

Material sources and data processing

Samples were collected in Xining City and Haibei Tibetan Autonomous Prefecture of Qinghai Province (Table 1). Samples were cleaned and stored at − 80 °C for later use. After leaf cleaning, DNA was extracted and sent to Nanjing Genesioneer Biotech Co., Ltd. for sequencing of chloroplast genomes of five *Caragana* species using the Illumina NovaSeq platform. sent.

After removing low-quality reads and adapter sequences, the chloroplast genome was assembled using SPAdes v3.10.1 (https://www.baseclear.com/services/bioinformatics/basetools/sspace-standard/). CDS sequences from the chloroplast genomes were then compared using Blast v2.2.25 (https://blast.ncbi.nlm.nih.gov/Blast.cgi). After manual correction, the annotation results were finalized. Annotation of rRNA and tRNA sequence was obtained using hmmerv3.1b2 (http://www.hmmer.org/) and ARAGORN v1.2.38 (http://130.235.244.92/ARAGORN/). The chloroplast genome data for the five *Caragana* species were uploaded to the NCBI database.

Sequence data filtering

To ensure the accuracy of the codon bias analysis for five *Caragana* species, further screening of the compiled and annotated data was performed. Seventy-six sequences were filtered using three criteria: (1) removal of duplicate sequences and those shorter than 300 base pairs; (2) retention of sequences with ATG as a start codon; (3) Selection of sequences with TAA, TGA or TAG as a stop codon. The 52 coding sequences (CDS) that met these criteria were saved in separate files in FASTA format for subsequent codon bias analysis. Table 2 presents the basic information of the chloroplast genomes for the five *Caragana* species after screening, while Table 3 lists the detailed names of the 52 CDS sequences that meet the criteria.

Experimental sequence data analysis methods

Codon usage indices analysis

Codon usage indices include GC content at the third position of synonymous codons (GC3s), codon adaptation index (CAI), optimal codon frequency (FOP), effective number of codons (ENC), relative synonymous codon Usage (RSCU) and the total GC content in coding sequences (CDS).

Table 2 Information of chloroplast genome of 5 *Caragana* plants

Species	The full sequence length / bp	Number of coding sequences	Accession number
Caragana arborescens	129 473	52	MT211962
C. jubata	128 132	52	MT211963
C. opulens	132 815	52	OQ656872
C. tibetica	128 433	52	OQ942026
C. turkestanica	129 453	52	OQ942027

Table 1 The satistical table of sample source and sequencing data

Species	Collection site	Total reads of Clean Data	GC content of Clean Data	Number of mapped pair-end reads
Caragana arborescens	Xining City, Qinghai Province (N36°43′24.80″, E101°44′54.11″)	22 855 947	38.31	1 063 570
Caragana jubata	Haibei Tibetan Autonomous Prefecture, Qinghai Province (N37°36′53.34″, E101°19′18.63″)	22 519 738	39.30	2 033 829
Caragana opulens	Xining City, Qinghai Province (N37°36′53.34″, E101°19′18.63″)	21 130 518	38.11	940 760
Caragana tibetica	Xining City, Qinghai Province (N36°43′24.80″, E101°44′54.11″)	17 696 991	38.22	2 100 287
Caragana turkestanica	Xining City, Qinghai Province (N36°43′24.81″, E101°44′54.12″)	22 676 575	37.45	1 577 272

Table 3 Fifty-two protein-coding genes in chloroplast genomes of five *Caragana* species

Gene classification	Gene grouping	Gene name
Self-replication related genes	RNA polymerase subunit gene	*rpoA* 、 *rpoB* 、 *rpoC1* 、 *rpoC2*
	NADH dehydrogenase gene	*ndhA* 、 *ndhB* 、 *ndhC* 、 *ndhD* 、 *ndhE* 、 *ndhF* 、 *ndhG* 、 *ndhH* 、 *ndhI* 、 *ndhJ* 、 *ndhK*
	ribulose bisphosphate carboxylase large subunit	*rbcL*
	ribosomal small subunit gene	*rps2* 、 *rps3* 、 *rps4* 、 *rps7* 、 *rps8* 、 *rps11* 、 *rps12* 、 *rps14* 、 *rps18*
Photosynthesis genes	Subunit of photosesten I	*psaA* 、 *psaB*
	Subunit of ATP synthase	*atpA* 、 *atpB* 、 *atpE* 、 *atpF* 、 *atpI*
	Cytochrome Complex	*petA* 、 *petB* 、 *petD*
	Subunit of photosesten II	*psbA* 、 *psbB* 、 *psbC* 、 *psbD*
Other genes	Envelop membrane protein	*cemA*
	c-type cytochrome synthesis	*ccsA*
	Maturase K	*matK*
	Subunit of Acetyl-CoA-Carboxyase	*accD*
	Proteases gene	*clpP*
	Large subunit of ribosome	*rpl2* 、 *rpl14* 、 *rpl16* 、 *rpl20*
	Conserved open reading frames	*ycf1* 、 *ycf2* 、 *ycf3* 、 *ycf4*

The online tool EMBOSS-cusp (https://www.bioinformatics.nl/cgi-bin/emboss/cusp) was used to calculate the total GC content (GCall) and the GC content at the first, second and third Calculate position (GC1, GC2, GC3) for 52 CDS sequences from five *Caragana* species. CodonW 1.4.2 (http://downloads.fyxm.net/CodonW-76666.html) was used to calculate and analyze RSCU, ENC, CAI, FOP and other related parameters for the 52 protein-coding genes of these five *Caragana* species that meet the specified criteria.

Neutral plot analysis

The GC1, GC2, and GC3 values obtained from EMBOSS-cusp were used to calculate the average GC content at the first and second positions (GC12) using Excel. The ggplot2 package in R was used to create a scatterplot with GC3 on the horizontal axis and GC12 on the vertical axis, including a best-fit line and a 95% confidence interval. This chart is commonly used to evaluate factors that affect codon usage patterns [22].

ENC-plot analysis

The GC3s and ENC values calculated with CodonW were plotted in a scatterplot, with GC3s serving as the horizontal axis and ENC as the vertical axis. A standard curve representing ENC values has been incorporated into the graph. This diagram allows for the examination of codon usage biases within individual genes and examines the influence of base composition on these biases [23]. Each point in the graph represents a specific gene. If the data points in the scatterplot cluster primarily around the standard curve, this means that codon usage is predominantly influenced by mutations. On the contrary, significant deviations from the curve indicate a more important role of natural selection and other influencing factors [24].

The formula for calculating the standard curve is as follows [23]:

$$ENC_{exp} = 2 + GC3_s + \frac{29}{GC3_s^2 + (1 - GC3_s)^2} \quad (1)$$

The ENC value derived from the standard curve is called the expected ENC (ENCexp), while the ENC value obtained by CodonW is called the observed ENC (ENCobs). The ENC ratio, calculated using Frank Wright's (1990) formula, further highlights the differences in codon usage bias between different genes. If ENCexp and ENCobs share similarity, this suggests that codon usage biases in the chloroplast genome are associated with variations in GC3s and are primarily shaped by mutation rather than selection [25].

The calculation formula of ENC ratio is as follows [26]:

$$ENC_{ratio} = \frac{ENC_{exp} - ENC_{obs}}{ENC_{exp}} \quad (2)$$

PR2 – plot analysis

The mutation bias between AT and CG at the third codon position is analyzed by PR bias analysis (PR2 bias plot analysis) [27], which examines the composition of the third base in codons encoding amino acids. CodonW

was used to determine the A, T, C, and G content at this position, specifically A3s, T3s, C3s, and G3s. The AT bias (A3s/(A3s + T3s)) and the GC bias (G3s/(G3s + C3s)) were calculated. A PR2 plot with G3s/(G3s + C3s) on the horizontal axis and A3s/(A3s + T3s) on the vertical axis visually represents the degree and direction of gene distortion [28]. This plot makes it easier to infer codon usage trends and the Intensity of selection pressure [25].

Optimal codon selection

The amino acid coding of different species shows a preference for specific synonymous codons, reflecting inherent biases in codon usage. Identifying optimal codons can improve translation efficiency and gene expression accuracy because highly expressed genes often have different codon usage patterns [29, 30]. Based on the actual ENC values of the sample genes, the 52 CDS sequences were sorted in descending order. Genes with ENC values in the top 10% and bottom 10% were then selected, representing the 5 genes with the highest and lowest ENC values, respectively. These genes were used to create gene libraries representing high and low codon preferences. RSCU values calculated by CodonW were categorized based on the libraries and the ΔRSCU was calculated between the high and low codon preference libraries [31]. The resulting RSCU values were visualized using R as a circular heatmap.

If the RSCU (Relative Synonymous Codon Usage) value of a codon exceeds 1, it indicates a higher frequency of usage. Codons with an ΔRSCU value greater than 0.08 are considered to be preferentially used in highly expressed genes. Codons that meet both criteria are classified as optimal codons [32].

Phylogenetic analysis

The RSCU values of five *Caragana* species were used to calculate a distance matrix in R for cluster analysis and dendrogram construction. Bayesian trees were constructed using PhyloSuite software, based on the chloroplast genome and coding DNA sequences (CDS) of these five *Caragana* species and other species.

Results

Analysis of codon bias-related parameters

Analysis of the CDS sequences of five *Caragana* species revealed total GCall percentages of 36.92%, 36.93%, 37.03%, 37.05% and 36.91% for *Caragana arborescens*, *Caragana jubata*, *Caragana opulens*, *Caragana tibetica* and *Caragana turkestanica*. The GC content at positions 1, 2, and 3 (GC1, GC2, GC3) was below 46% in all species, with GC1 consistently higher than GC2 and GC3, indicating a trend of GC1 > GC2 > GC3 (Table 4).

Table 4 GC content of the codon of chloroplast genomes in five *Caragana* species

Species	GC1/%	GC2/%	GC3/%	GCall/%
C. arborescens	45.36	37.42	27.97	36.92
C. jubata	45.41	37.44	27.94	36.93
C. opulens	45.50	37.39	28.20	37.03
C. tibetica	45.49	37.42	28.24	37.05
C. turkestanica	45.35	37.41	27.98	36.91

The codon parameter properties include the GC content at different codon positions as well as the effective number of codons (ENC), the codon adaptation index (CAI), the codon bias index (CBI), and the optimal codon frequency (Fop). The effective number of codons (ENC) is crucial for evaluating the codon uses bias. Lower values indicate a preference for rare codons and a threshold of 35 is commonly used to distinguish between strong and weak bias [30]. The ENC value ranges from 20 to 61, with values closer to 20 indicating stronger bias [22], and values closer to 61 indicating weaker bias [33].

The Codon Adaptation Index (CAI) assesses the similarity of codon usage in genes to that of highly expressed genes and serves as a tool to assess gene expression levels, with higher CAI values indicating higher expression [34]. The Codon Bias Index (CBI) assesses gene expression by calculating the proportion of optimal codons used in a gene, where a value of 1 indicates full utilization of optimal codons and negative values indicate underutilization. The optimal codon frequency (Fop) ranges from 0 (no optimal codons used) to 1 (full use of optimal codons) [35].

The parameters calculated by CodonW for the five *Caragana* species are shown in Fig. 1. The ENC values of all five *Caragana* species fell between 35 and 60, indicating a weak codon bias. CAI values ranging from 0 to 1 indicate a stronger codon usage preference at higher values, implying higher expression levels [36]. CBI values, which are mostly negative and between -0.5 and 0.25, indicate suboptimal use of optimal codons. The Fop values indicated variable but generally low frequencies of optimal codon usage among the five *Caragana* species.

Correlation coefficients between parameters related to codon usage bias for the five *Caragana* species were calculated using SPSS and a correlation heatmap was created. The heatmap shows that the correlation coefficients between GC3s and GC3 are 0.928 for *Caragana arborescens* and 0.938 for *Caragana jubata*, indicating a highly significant correlation. In contrast, the correlation coefficients for *Caragana opulens*, *Caragana tibetica* and *Caragana turkestanica* are 0.158, 0.177 and 0.198, respectively, showing no significant correlation. ENC has

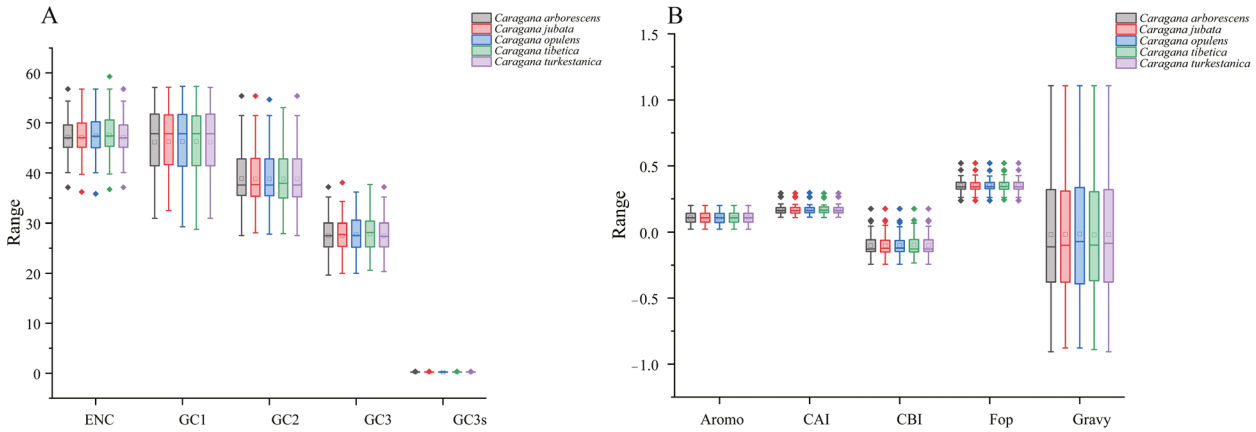

Fig. 1 Codon usage bias related parameters of five *Caragana* species

a positive correlation with GC1 in *Caragana arborescens* and *Caragana jubata*, but a negative correlation in *Caragana opulens*, *Caragana tibetica* and *Caragana turkestanica*. The correlation coefficient between CAI and GC1 is significant in *Caragana arborescens* and *Caragana jubata*, but not in the other three species (Fig. 2).

Neutrality plot analysis

Scatterplots were created for each of the five *Caragana* species. The red line represents the best-fit curve, while the blue shaded area represents the 95% confidence interval of that curve. The slope of the best-fit curve correlates with the base composition at different codon positions, allowing conclusions to be drawn about the factors that influence codon bias. A steeper slope indicates greater similarity in base composition, meaning that codon bias is caused primarily by internal mutations. Conversely, a flatter slope indicates a greater influence of natural selection [35]. The GC3 values are between 19 and 39% and the GC12 values between 30 and 59%. The regression coefficients vary between 0.35 (*Caragana opulens*) and 0.43 (*Caragana tibetica*),

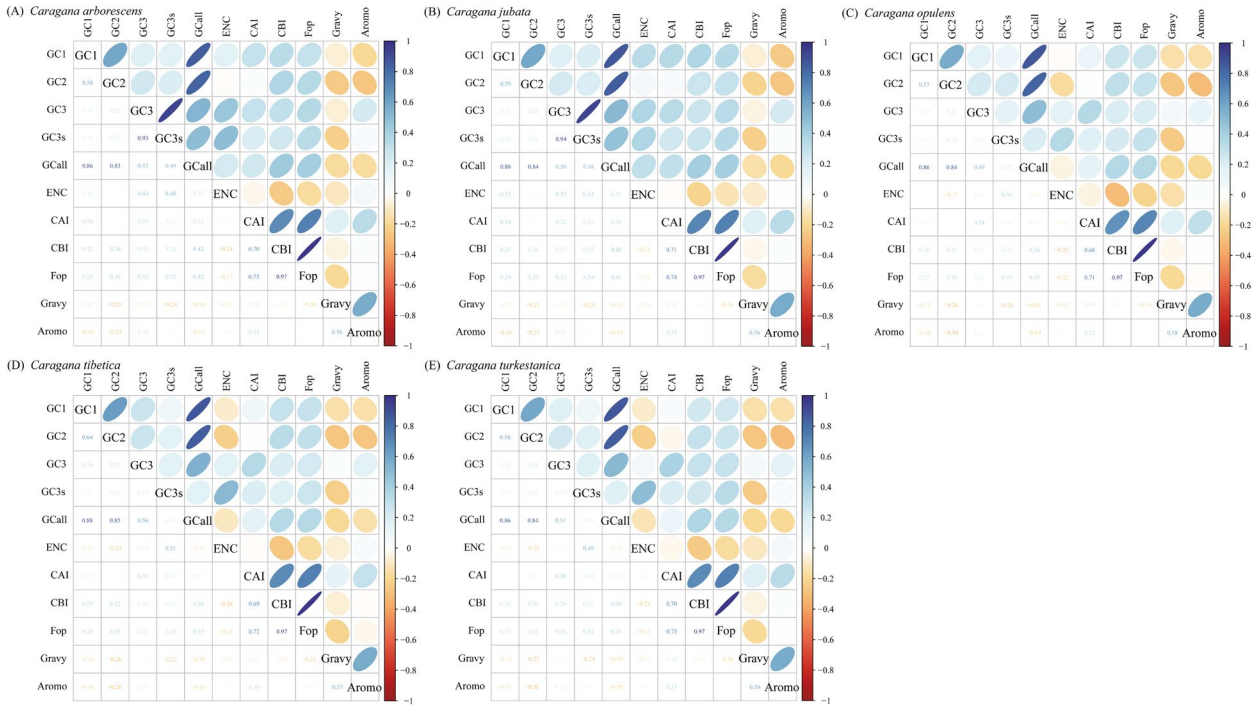

Fig. 2 Correlation analysis between parameters of each gene of five *Caragana* species

indicating that internal mutations contribute up to 43% to the codon bias. This suggests that natural selection plays a more important role than mutations in shaping the codon bias between the five *Caragana* species (Fig. 3).

ENC-plot analysis

Scatter plots were constructed using the ENC and GC3s values of the five *Caragana* species and an ENC expected value curve was included to assess the effect of mutations on codon bias (Fig. 4). The broad distribution of scores across different genes indicates significant differences in codon bias.

ENC ratios were calculated using the formulas ① and ②, and stacked bar graphs (Fig. 5) were created to illustrate the distribution of ENC ratios among the five *Caragana* species. Most genes are in the (0.00, 0.10] interval, followed by the (-0.10, 0.00] interval, with the fewest genes in the (0.20, 0.30] interval, which is completely absent in *Caragana tibetica.* The narrow range of ENC deviations suggests that the codon bias in these species is mainly caused by mutations.

PR2-plot analysis

Considering only base mutations, random mutations would result in equal probabilities of A/T and C/G at the third codon position. However, if natural selection affects codon bias, it would result in unbalanced usage of A/T and G/C [37].

PR2 plot analysis of the *Caragana arborescens* chloroplast genome revealed that 31 genes had a G3s/(G3s + C3s) ratio greater than 0.5, while 35 genes had an A3s/(A3s + T3s) ratio less than had 0.5. No gene showed a central point of 0.5, indicating that G occurred more frequently than C at the third codon position and T occurred more frequently than A. Similarly, in the chloroplast genome of *Caragana jubata*, 31 genes had a G3s/(G3s + C3s)- Ratio > 0.5 and 34 genes had A3s/(A3s + T3s) ratio < 0.5, confirming the preference of G over C and T over A at the third codon position. In *Caragana opulens*, 32 genes had G3s/(G3s + C3s) ratio > 0.5, while 36 genes had A3s/(A3s + T3s) ratio < 0.5, showing a similar trend. In *Caragana tibetica*, 33 genes had a G3s/(G3s + C3s) ratio > 0.5, while 36 genes had an A3s/(A3s + T3s) ratio < 0.5, reinforcing the overall pattern. Finally, in *Caragana turkestanica*, 31 genes had a G3s/(G3s + C3s) ratio > 0.5, while 35 genes had an A3s/(A3s + T3s) ratio < 0.5, with no central point at 0.5, which indicating a consistent preference for G and T over C and A at the third codon position. These results suggest that the bias in codon usage in the chloroplast genomes of the five *Caragana* species is influenced by natural selection (Fig. 6).

Optimal codon analysis

As shown in Fig. 7, by setting the RSCU (Relative Synonymous Codon Usage) threshold to more than 1, 29 highly frequent codons were identified in all five *Caragana species*. Of these codons, 55.2% ended in U, 41.4% in A, and only one codon (UUG, encoding leucine) ended in G. This further confirms the preference for A/U-ending

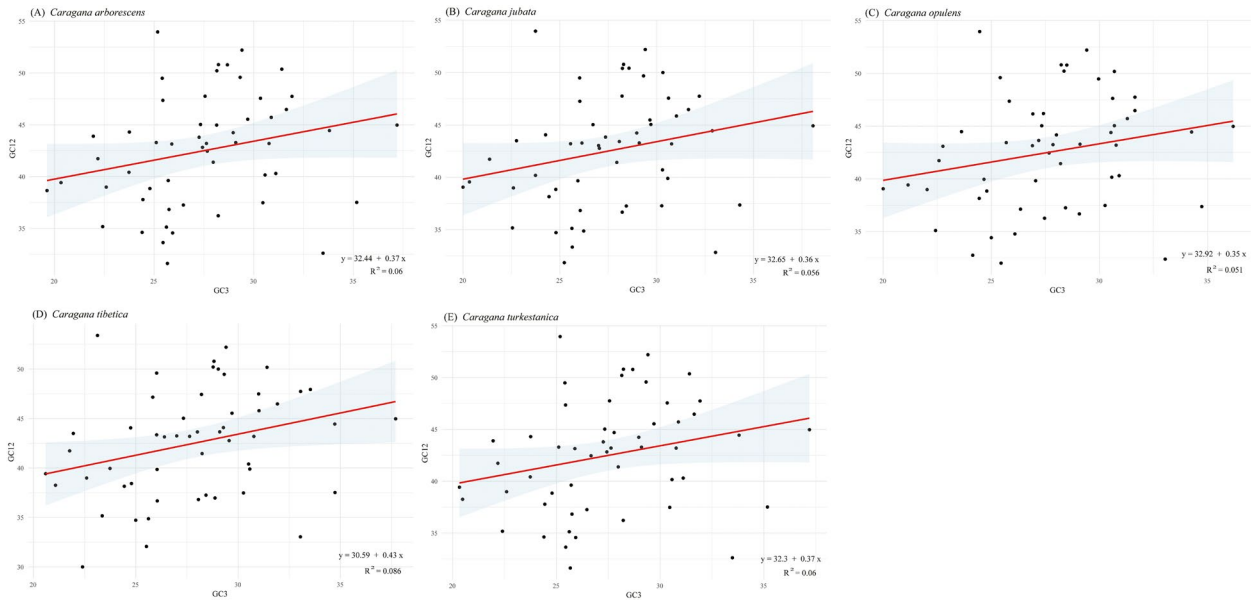

Fig. 3 Neutrality plot analysis of five *Caragana* species

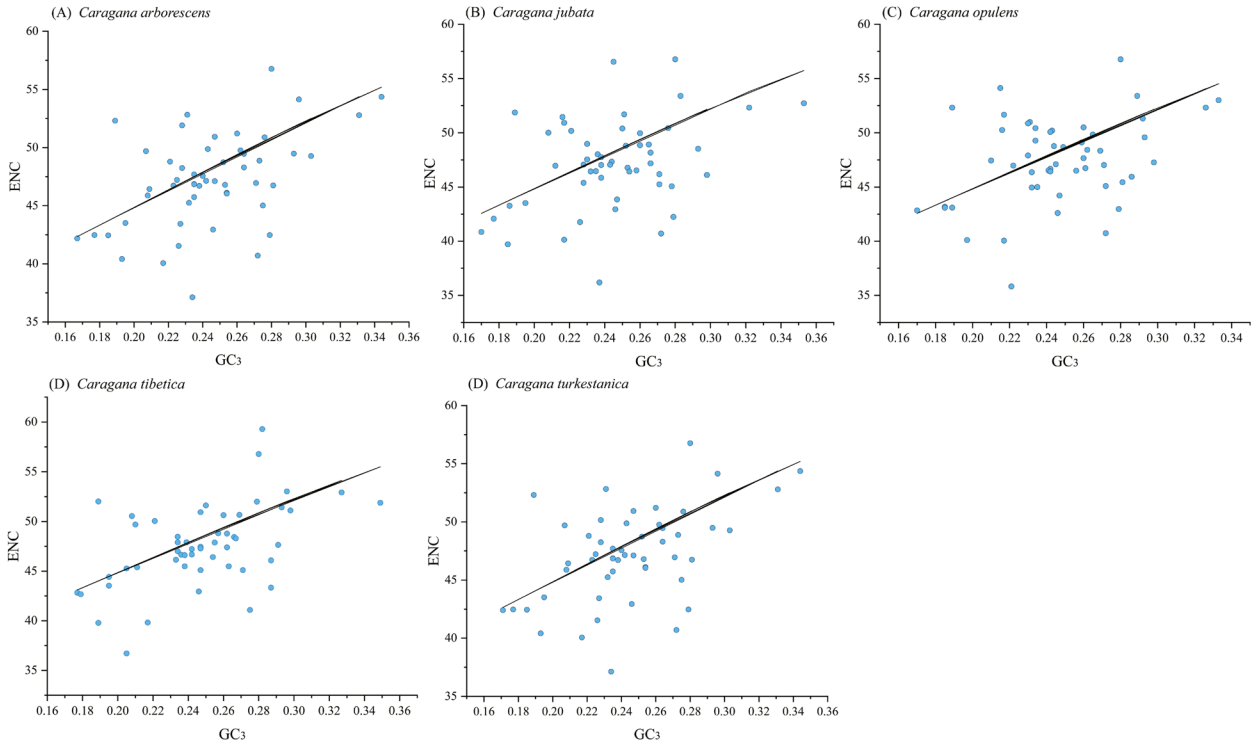

Fig. 4 Neutrality plot analysis of five *Caragana* species

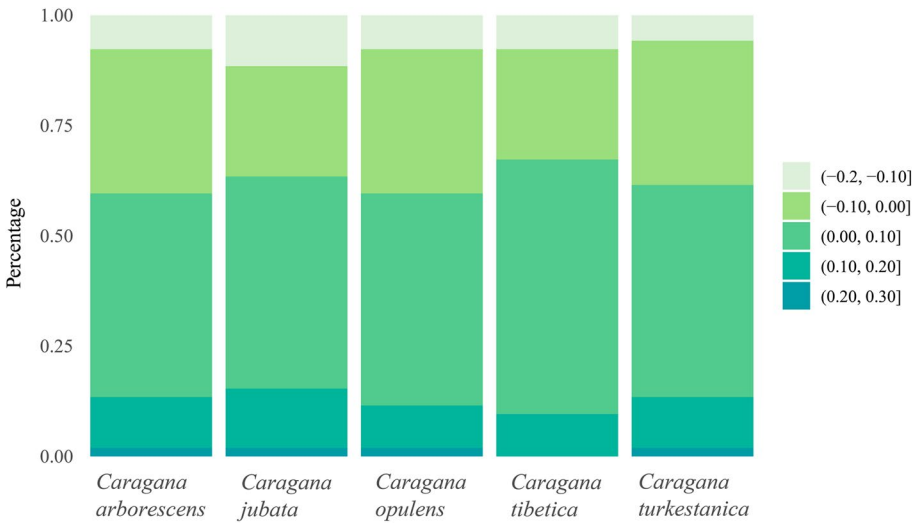

Fig. 5 ENC ratio frequency distribution plot of five *Caragana* species

codons in *Caragana* chloroplast genes. Highly expressed optimal codons were selected by applying an additional RSCU criterion (the difference between the RSCU of a codon and the average RSCU for the same amino acid) of at least 0.08.

Caragana tibetica, *Caragana turkestanica*, *Caragana arborescens*, *Caragana opulens*, and *Caragana jubata* had 26, 31, 31, 26, and 28 highly expressed optimal codons, respectively. Effective number of codons (ENC) values were used to categorize genes by expression levels, and optimal codons in the chloroplast

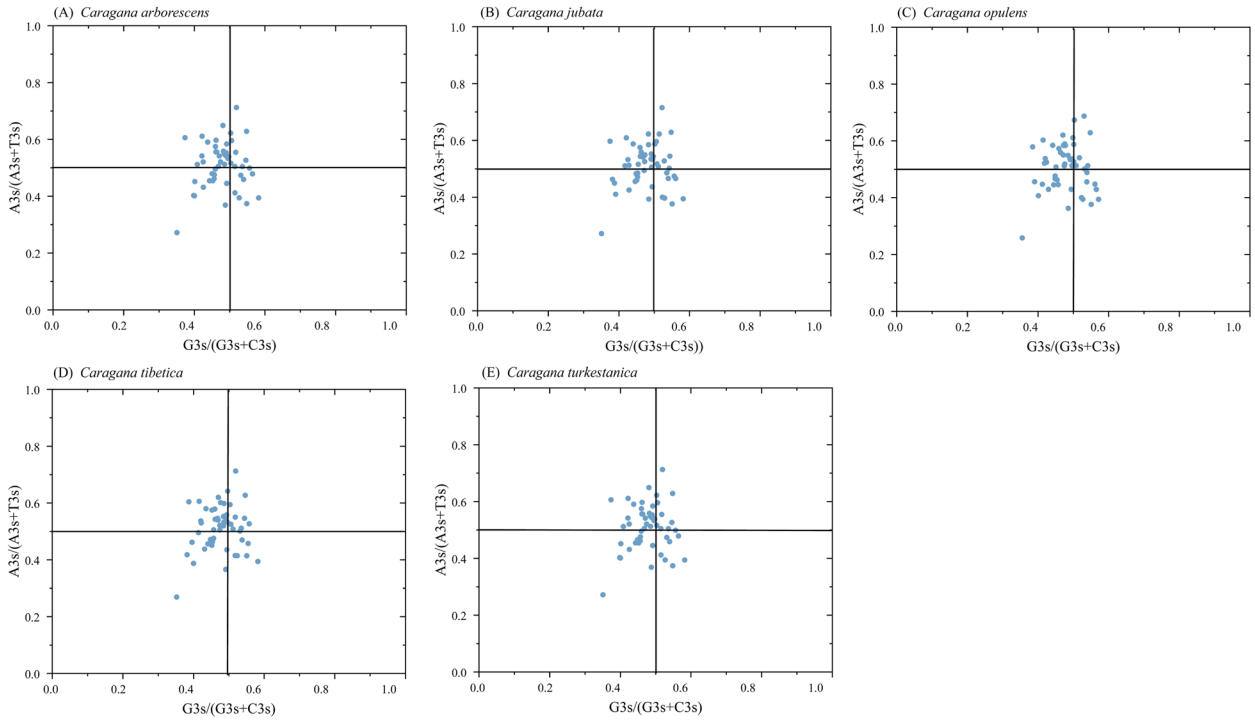

Fig. 6 PR2-plot of five *Caragana* species

genomes were identified based on RSCU > 1 and ΔRSCU ≥ 0.08. The final selection identified 10, 9, 9, 12, and 11 optimal codons for *Caragana tibetica*, *Caragana turkestanica*, *Caragana arborescens*, *Caragana opulens*, and *Caragana jubata*, respectively.

Phylogenetic analysis

Clustering of the five *Caragana* species based on RSCU values showed that *Caragana tibetica*, *Caragana opulens*, and *Caragana jubata* formed separate groups, whereas *Caragana arborescens* and *Caragana turkestanica* clustered together (Fig. 8A). This suggests similar codon preferences between the latter two species and distinct preferences among the others. Phylogenetic trees based on chloroplast genome CDS sequences (Fig. 8B) and complete chloroplast genome sequences (Fig. 8C) indicated that *Caragana korshinskii*, *Caragana kozlowii* and *Caragana microphylla* cluster together in one branch. *Caragana jubata* and *Caragana tibetica* have the closest phylogenetic relationship, while *Caragana opulens* and *Caragana rosea* var. *rosea* also have a relatively close phylogenetic relationship, and these two sub-branches cluster together into a larger branch. In contrast, *Caragana arborescens* and *Caragana turkestanica* have a more distant phylogenetic relationship with other *Caragana* species.

Discussion

In the above study, the codon use bias parameters showed an effective number of codons (ENC) ranging from 35.82 to 56.77, CAI and FOP indices near zero, and all CBI indices negative. These results suggest relatively low levels of chloroplast gene expression in the five Caragana species. Similar patterns have been observed in other angiosperms, such as *Lonicera* species [38] and *Glycyrrhiza* species [39].

We conducted an in-depth investigation of the factors influencing codon bias in the chloroplast genomes of *Caragana* species. ENC plot analysis revealed that natural selection predominantly influences most genes, whereas a smaller proportion is influenced by mutation effects. Further PR2 plot analysis revealed that mutation pressure has no influence on codon bias. Rather, natural selection predominantly shapes the codon usage patterns during the evolution of the five *Caragana* species, with gene mutations having no significant influence. This

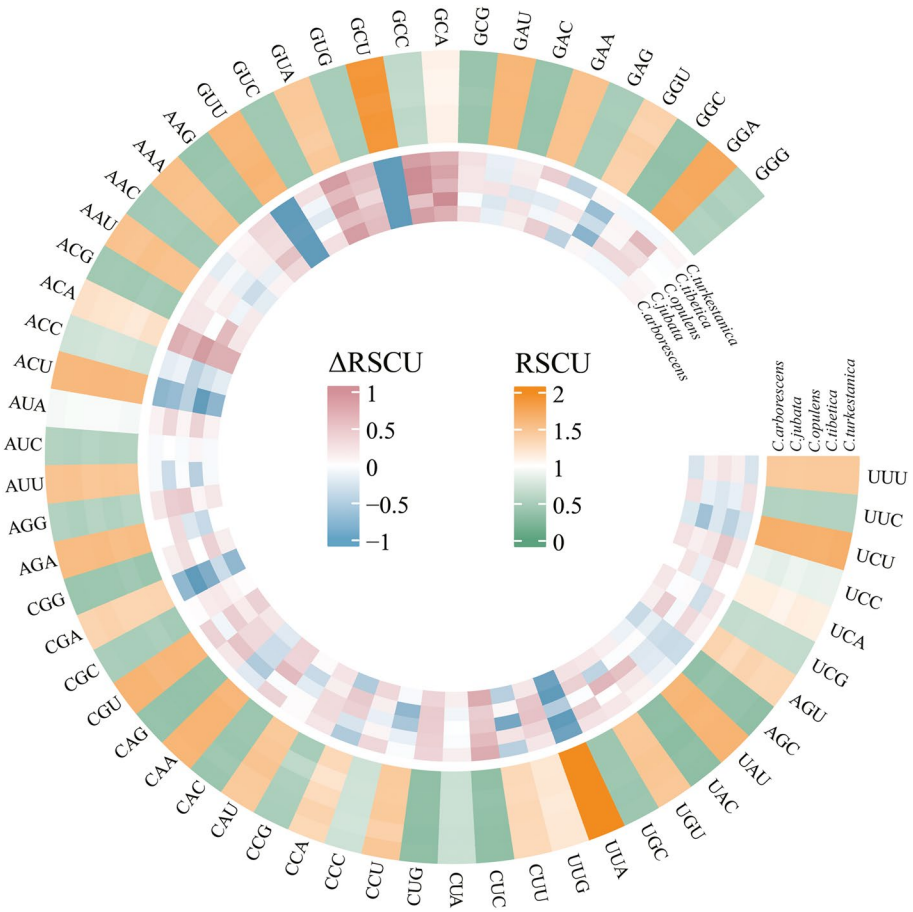

Fig. 7 Relative synonymous codon usage and optimal codon analysis of five *Caragana* species

finding highlights the critical role of natural selection in shaping *Caragana* chloroplast gene evolution.

This process is inextricably linked to the biological functions of chloroplasts, which serve as primary organelles for photosynthesis in plants and convert solar energy into chemical energy [40]. During photosynthesis, chloroplasts provide *Caragana* plants with energy and vital organic molecules, supporting their normal growth and development. Studies on other plants such as *Elaeagnus angustifolia* [41] and *Nymphaea tetragona* [42] have also shown the influence of natural selection on the evolution of chloroplast genes, similar to observations in *Caragana*.

By setting an RSCU threshold greater than 1, we were able to identify 29 high-frequency codons common to all five *Caragana* species: *Caragana tibetica*, *Caragana turkestanica*, *Caragana arborescens*, *Caragana opulens* and *Caragana jubata*. Our research revealed a strong preference for codons ending in A or U in the *Caragana* chloroplast genes. The research found that codon usage in the chloroplast genes of Caragana species shows a significant bias towards codons ending in A or U [12]. This phenomenon is consistent with the theory proposed by Morton, which states that genomes with low GC content tend to prefer codons ending in A or U. This A/U bias in the chloroplast genomes of plants due to the lack of inverted repeat regions (IRs) may be related to the simplification and functional optimization of their genome structure [21].

By applying a more stringent criterion of RSCU 0.08 to select highly expressed optimal codons and combining this with ENC values to create a gene library based on expression levels, we have created the conditions for the identification of optimal codons.

This approach identified 10, 9, 9, 12, and 11 optimal codons for the five *Caragana* species, respectively, reinforcing the preference for A/U-ending codons in their chloroplast genes. Yu Xiao et al. [43] found that in their study of codon bias in the chloroplast genomes of *Amphicarpaea ferruginea* and *Amphicarpaea edgeworthii*, the optimal codons of both species ended in A/U.

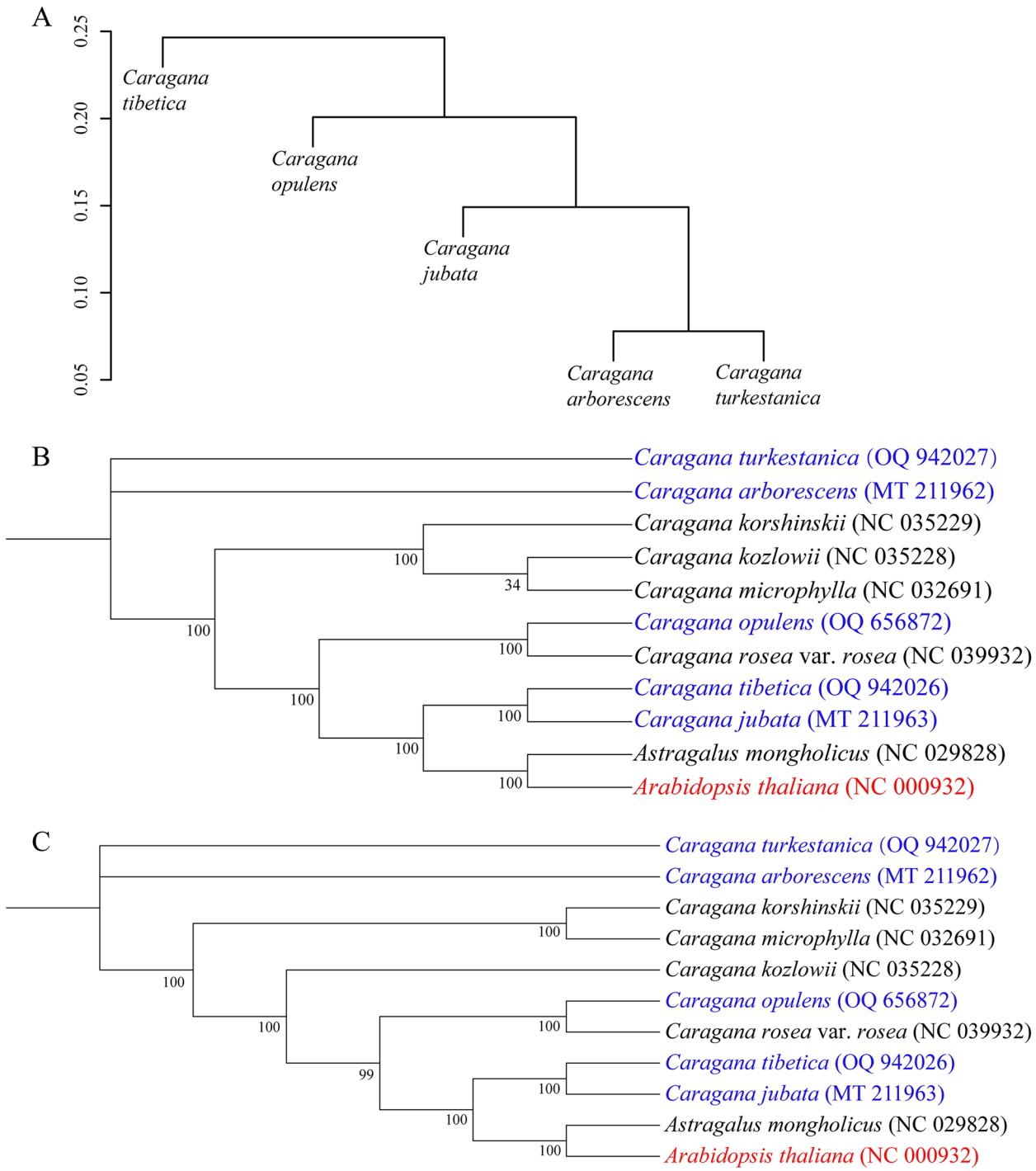

Fig. 8 Phylogenetic tree of five *Caragana* species. **A** RSCU-based clustering analysis dendrogram; **B** phylogenetic tree based on CDS sequences of chloroplast genome; **C** phylogenetic tree based on the whole sequence of chloroplast genome

The preference for A/U-ending codons in the chloroplast genes of *Caragana* species is a well-documented phenomenon in plant molecular evolution. This preference is supported by several theoretical and empirical studies that highlight the role of codon bias in shaping gene expression and evolutionary adaptation [12].

The preference for A/U-ending codons likely enhances translation efficiency and accuracy. This is particularly

relevant for chloroplast genes, which are essential for photosynthesis and other metabolic processes. Efficient translation is crucial for maintaining high levels of protein synthesis, especially under environmental stress conditions [44]. The consistent preference for A/U-ending codons across multiple *Caragana* species indicates a high degree of conservation in the chloroplast genomes. This conservation is likely driven by functional constraints and natural selection favoring codons that enhance translation efficiency and accuracy [45].

This suggests that the preference for A/U-ending codons is a common feature in plant chloroplast genomes, potentially driven by similar selective pressures.In summary, the preference for A/U-ending codons in the chloroplast genes of *Caragana* species is supported by well-established theories in molecular evolution. This preference likely enhances translation efficiency and accuracy, contributing to the ecological adaptability of these species. The ecological adaptability of *Caragana* species to harsh environments, such as drought and extreme temperatures, may be partly attributed to these genetic adaptations.

The RSCU values clustering analysis provides a quantitative measure of codon usage similarity among species. By clustering the RSCU values, we can identify species with similar codon usage patterns, which often reflects their evolutionary relatedness [11]. The phylogenetic trees constructed from CDS and whole-genome sequences provide a visual representation of the evolutionary relationships among the species.

To provide a comprehensive understanding of the evolutionary relationships among the five *Caragana* species, we performed clustering analysis based on RSCU values and constructed phylogenetic trees using both CDS and whole-genome sequences. The clustering analysis of RSCU values revealed distinct groupings among the species, with *Caragana arborescens* and *Caragana turkestanica* forming a closely related cluster These results are also consistent with the sub-classification of *Caragana arborescens* and *Caragana turkestanica* in the Flora of China (https://www.iplant.cn/frps). They all belong to Ser. *Caragana*. This clustering was further supported by the phylogenetic trees constructed from both CDS and whole-genome sequences, indicating a strong evolutionary relationship between these two species [2, 46].

Similarly, *Caragana tibetica* and *Caragana jubata* were found to cluster together in both the CDS and whole-genome phylogenetic trees, suggesting a close evolutionary relationship between these species as well [13]. In contrast, *Caragana opulens* was found to be relatively distant from *Caragana arborescens* in both the RSCU clustering and phylogenetic analyses, indicating a more distant evolutionary relationship [14]. These results are consistent with the study by Lie Liu et al. [20, 47] on the phylogenetic relationships of chloroplast genomes in four *Caragana species*. The results from the three types of phylogenetic trees—based on RSCU values, CDS sequences, and whole-genome sequences—complement each other, providing a more robust and authentic representation of the evolutionary relationships within the *Caragana* species.

This study not only improved the understanding of codon usage bias, but also provided insights into the similarities and differences in the regulation of chloroplast gene expression between *Caragana* species and other species in Leguminosae.

Conclusion

The GC content is similar in *Caragana arborescens*, *Caragana jubata* and *Caragana turkestanica*, while *Caragana opulens* and *Caragana tibetica* also have similar GC content. All five *Caragana* species prefer codons ending in A/U. The codon bias is relatively weak in these *Caragana* species because optimal codons are rarely used. Natural selection influences the codon usage patterns of these *Caragana* species significantly more than mutations. There is a remarkable correlation between some parameters related to codon usage bias in *Caragana arborescens* and *Caragana jubata*. Phylogenetic analysis shows that *Caragana tibetica* and *Caragana jubata* are sister species, forming a sister group with *Caragana opulens*, while *Caragana arborescens* forms a sister group with itself. Analysis of codon bias in the chloroplast genomes of *Caragana* species is crucial for understanding codon usage preferences and the molecular mechanisms underlying their environmental adaptation.

Acknowledgements
We would like to thank Na Hu and Jing Sun for their help in researching the materials.

Statement
Our experimental research and field studies on plants comply with relevant institutional, national, and international guidelines and legislation.

Authors' contributions
XL designed and executed experiments, completed data analysis, and wrote the first draft of the paper. LL, QR, and TZ contributed to the experimental design and analysis. NH and JS assisted in sample collection and species identification. WZ 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
National Natural Science Foundation (No. 32160386) of China and The Open Project of Qinghai Key Laboratory of Qinghai-Tibet Plateau Biological Resources (2024-KF-04).

Data availability
The original sequencing data have been submitted to the NCBI database and received GenBank accession numbers OQ942026 (C. tibetica), OQ942027 (C. turkestanica), MT211962 (C. arborescens), OQ656872 (C. opulens), MT211963

(C. jubata). The data used in this study are already entirely in the public domain (https://www.ncbi.nlm.nih.gov). Voucher specimens of C. tibetica, C. turkestanica, C. arborescens, C. opulens and C. jubata are stored in the herbarium of the School of Ecological and Environmental Engineering at Qinghai University. The voucher specimen number for Caragana tibetica is QhST20190080, for Caragana arborescens is QhST20190078, for Caragana jubata is QhST20190077, for Caragana opulens is QhST20190079 and for Caragana turkestanica is QhST20190081.

Declarations

Ethics approval and consent to participate

C. tibetica, *C. turkestanica, C. arborescens, C. opulens* and *C. jubata* were collected in September 2019 from non-private land, and anyone is permitted to collect these wild plants for research purposes without causing ecological harm. Voucher specimens of *C. tibetica*, *C. turkestanica, C. arborescens, C. opulens* and *C. jubata* are stored in the herbarium of the School of Ecological and Environmental Engineering at Qinghai University. The botanical identification was performed by the corresponding author, Dr. Zhou. The voucher specimen number for *Caragana tibetica* is QhST20190080, for *Caragana arborescens* is QhST20190078, for *Caragana jubata* is QhST20190077, for *Caragana opulens* is QhST20190079 and for *Caragana turkestanica* is QhST20190081.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 29 September 2024 Accepted: 4 March 2025
Published online: 13 March 2025

References

1. Sabir J, Schwarz E, Ellison N, et al. Evolutionary and biotechnological implications of plastid genome variation in the inverted-repeat-lacking clade of legumes. Plant Biotechnology J. 2014;12(6):743–54.
2. Meng Q, Niu Y, Niu X, et al. Ethnobotany, phytochemistry and pharmacology of the genus Caragana used in traditional Chinese medicine. J Ethnopharmacology. 2009;124(3):350–68.
3. Moukoumi J, Hynes RK, Dumonceaux TJ, et al. Characterization and genus identification of rhizobial symbionts from Caragana arborescens in western Canada. C J of Microbiol. 2013;59(6):399–406.
4. Kordyum E, Bilyavska N. Structure and biogenesis of ribonucleoprotein bodies in epidermal cells of Caragana arborescens L. Protoplasma. 2018;255(2):709–13.
5. Wang L, Yang X, Zhang Y, et al. Anti-inflammatory chalcone–isoflavone dimers and chalcone dimers from Caragana jubata. J Nat Prod. 2019;82(10):2761–7.
6. Ma C, Gao Y, Li Q, et al. Water regulation characteristics and stress resistance of Caragana opulens population in different habitats of Inner Mongolia plateau. J Appl Ecol. 2006;17(2):187–91.
7. Wu Z. Flora of China. Beijing, China: Science Press; 1993. p. 32.
8. Zhang M. Morphological Features of Caragana tibetica nebkhas in Mu Us Desert. J Changchun Normal Univ. 2019;38(12):72–8.
9. Zhang P, Yang J, Zhao L, et al. Effect of Caragana tibetica nebkhas on sand entrapment and fertile islands in steppe-desert ecotones on the Inner Mongolia Plateau. China Plant Soil. 2011;347:79–90.
10. Wu Z. Flora of China. Beijing, China: Science Press; 1993. p. 42.
11. Sharp PM, Li WH. The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 1987;15(3):1281–95.
12. Morton BR. Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages. J Mol Evol. 1998;46(4):449–59.
13. Zhao C, Peng L, Wang X, et al. Codon bias and evolution analysis of AtGAI in Amaranthus tricolor L. Journal of China Agricultural University. 2019;24(12):10–22.
14. UDDIN A. Codon usage bias: a tool for understanding molecular evolution. Proteomics and Bioinformatics. 2017;10(5):10.4172.
15. Sang J, Wang Y, Li X, et al. Analysis of codon bias in WRKY transcription factors related to seawater stress in Hemerocallis fulva. J Technol. 2024;24(02):245–53.
16. Ghaemmaghami S, Huh WK, Bower K, et al. Global analysis of protein expression in yeast. Nature. 2003;425(6959):737–41.
17. Campos JL, Zeng K, Parker DJ, et al. Codon usage bias and effective population sizes on the X chromosome versus the autosomes in Drosophila melanogaster. Mol Biol Evol. 2013;30(4):811–23.
18. Chen Y, Shi Y, Deng H, et al. Characterization of the porcine epidemic diarrhea virus codon usage bias. Infect Genet Evol. 2014;28:95–100.
19. Wicke S, Schneeweiss GM, dePamphilis CW, Müller KF, Quandt D. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol. 2011;76(3–5):273–97.
20. Liu L, Li H, Li J, et al. Chloroplast genomes of Caragana tibetica and Caragana turkestanica: structures and comparative analysis. BMC Plant Biol. 2024;24(1):254.
21. Jansen RK, Ruhlman TA. Plastid genomes of seed plants. In: Bock R, Knoop V, editors. Genomics of Chloroplasts and Mitochondria. Springer; 2012. p. 103–26.
22. Li G, Pan Z, He Y, et al. Analysis of codon bias of chloroplast genome in Porphyra yezoensis. Gen Appl Biol. 2020;39(12):5789–95.
23. Li J, Qin Z, Guo C, et al. Codon bias in the chloroplast genome of Gelidocalamus tessellatus. J Bamboo Res. 2019;38(02):79–87.
24. Yang L, Dong Z, Wang Y, et al. Analysis on codon usage bias of chloroplast genome in Potentilla glabra var. mandshurica. Mol Plant Breed. 2022;20(04):1095–103.
25. Yang G, Su K, Zhao Y, et al. Analysis of codon usage in the chloroplast genome of Medicago truncatula. Acta Pratacul Sin. 2015;24(12):171–9.
26. Wright F. The "effective number of codons" used in a gene. Gene. 1990;87(1):23–9.
27. Jiang W, Lv B, He J, et al. Codon usage bias in the straw mushroom Volvariella volvacea. Chin J Biotechnol. 2014;09:1424–35.
28. Sueoka N. Translation-coupled violation of Parity Rule 2 in human genes is not the cause of heterogeneity of the DNA G+C content of third codon position. Gene. 1999;238(1):53–8.
29. Zhou H, Yan B, Chen S, et al. Evolutionary characterization of Tembusu virus infection through identification of codon usage patterns. Infect Genet Evol. 2015;35:27–33.
30. He W, Zhang H, Zhang Y, et al. Codon usage bias in the N gene of rabies virus. Infect Genet Evol. 2017;54:458–65.
31. Tian C, Li X, Li C, et al. Genome-wide analysis of codon usage bias in Saccharum species and its phylogenetically related species Erianthus fulvus. Biotechnol Bull. 2024;40(03):202–14.
32. Feng Z, Jiang Y, Zheng Y, et al. Codon use bias analysis of chloroplast genome of Cistanche. Chinese Trad Herbal Drugs. 2023;05:1540–50.
33. Liu H, Wang M, Yue W, et al. Analysis of codon usage in the chloroplast genome of Broomcorn millet (Panicum miliaceum L.). Plant Sci J. 2017;03:362–71.
34. Xu C, Cai X, Qian B, et al. Codon usage bias in Vitis vinifera. Acta Botan Boreali-Occiden Sin. 2012;32(02):409–15.
35. Lavner Y, Kotlar D. Codon bias as a factor in regulating expression via translation rate in the human genome. Gene. 2005;345(1):127–38.
36. Li H, Ji X, Zhu R, et al. Codon usage bias analysis for Octahydrolycopene Synthase Gene (PSY) from ten plant species. Acta Agriculturae Boreali-occidentalis Sinica. 2020;29(02):276–84.
37. Du Y, Li X, Jia X, et al. Codon preference analysis of chloroplast genome of Chinese wolfberry. Chinese Trad Herbal Drugs. 2024;55(04):1316–25.
38. Wang J, Ma D, Han X, et al. Analysis on codon usage bias of chloroplast genomes of 24 Lonicera materials. J Plant Res Environ. 2023;32(03):12–23.
39. Dai J, Cai Y, Liu Q, et al. Analysis of codon usage bias of chloroplast genome in seven Glycyrrhiza species. Chinese Traditl Herbal Drugs. 2023;54(09):2907–16.
40. Du M, Wang W, Bao H, et al. Analysis of codon bias in chloroplast genome of Trigonella Foenum-graecum. Acta Agrestia Sinica. 2024;32(02):409–18.
41. Wang J, Wang T, Wang L, et al. Assembling and analysis of the whole chloroplast genome sequence of Elaeagnus angustifolia and its codon usage bias. Acta Botan Boreali-Occiden Sin. 2019;39(09):1559–72.

42. Mao L, Huang Q, Long L, et al. Comparative analysis of codon usage bias in chloroplast genomes of seven Nymphaea Species. J Northwest Forestry Univ. 2022;37(02):98–107.
43. Yu X, Zhao Z, Deng L. Genomic characteristics and codon usage bias of chloroplast genome in Amphicarpaea Elliot. Acta Agriculturae Boreali-occidentalis Sinica. 2024;33(03):435–51.
44. Kawabe A, Miyashita NT. Patterns of codon usage bias in three dicot and four monocot plant species. Genes Genet Syst. 2003;78(5):343–52.
45. Akashi H, Gojobori T. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. Proc Natl Acad Sci U S A. 2002;99(6):3695–700.
46. Wicke S, Müller KF, de Pamphilis CW, et al. Mechanisms of functional and physical genome reduction in photosynthetic and nonphotosynthetic parasitic plants of the broomrape family. Plant Cell. 2013;25(10):3711–25.
47. Liu L, Li H, Li J, et al. Chloroplast genome analyses of Caragana arborescens and Caragana opulens. BMC Genom Data. 2024;25(1):16.

Publisher's Note