RESEARCH ARTICLE

Codon usage bias and genetic diversity in chloroplast genomes of *Elaeagnus* species (Myrtiflorae: Elaeagnaceae)

Changle Li[1] · Ling Zhou[1] · Jiangbo Nie[1] · Songping Wu[1] · Wei Li[2] · Yonghong Liu[1] · Yulin Liu[1]

Abstract Codon usage bias (CUB) reveals the characteristics of species and can be utilized to understand their evolutionary relationship, increase the target genes' expression in the heterologous receptor plants, and further provide theoretic assistance for correlative study on molecular biology and genetic breeding. The chief aim of this work was to analyze the CUB in chloroplast (cp.) genes in nine *Elaeagnus* species to provide references for subsequent studies. The codons of *Elaeagnus* cp. genes preferred to end with A/T bases rather than with G/C bases. Most of the cp. genes were prone to mutation, while the *rps7* genes were identical in sequences. Natural selection was inferred to have a powerful impact on the CUB in *Elaeagnus* cp. genomes, and their CUB was extremely strong. In addition, the optimal codons were identified in the nine cp. genomes based on the relative synonymous codon usage (RSCU) values, and the optimal codon numbers were between 15 and 19. The clustering analyses based on RSCU were contrasted with the maximum likelihood (ML)-based phylogenetic tree derived from coding sequences, suggesting that the t-distributed Stochastic Neighbor Embedding clustering method was more appropriate for evolutionary relationship analysis than the complete linkage method. Moreover, the ML-based phylogenetic tree based on the conservative *matK* genes and the whole cp. genomes had visible differences, indicating that the sequences of specific cp. genes were profoundly affected by their surroundings. Following the clustering analysis, *Arabidopsis thaliana* was considered the optimal heterologous expression receptor plant for the *Elaeagnus* cp. genes.

Keywords Codon usage bias · Chloroplast genomes · *Elaeagnus* · Optimal codons · Evolutionary relationship

Introduction

Codons, as the link between deoxyribonucleic acids and proteins in organisms, play an essential role in transmitting genetic information (Liu et al. 2020). Codon usage bias (CUB) refers to the non-random use of synonymous codons to encode the amino acids of a protein in the deoxyribonucleic acid transcription. Prior studies have shown that CUB was universal and confirmed in many organisms, such as bacteria, fungi, animals, and plants. CUB differs not only among species but also within the cell, such as the nucleus, chloroplast (cp.), and mitochondrion of a cell, and even between various genes of the same genome (Liu et al. 2017). Accompanying the swift progress of high-throughput sequencing techniques in recent years, genomic sequencing of multiple species has been accomplished, helping to understand the CUB at the genome-wide level. The CUB in gene families and the whole genome has been intensively studied in many model and non-model organisms, such as *Arabidopsis thaliana* (Chiapello et al. 1998), *Nicotiana tabacum* (Anwar et al. 2021), phytoplankton (Krasovec and Filatov 2022), and *Gnetum luofuense* (Deng et al. 2021b). By CUB analysis, genetic and mutational events influencing

✉ Yonghong Liu
lyhlxy2006@163.com

✉ Yulin Liu
liuyulin@nwafu.edu.cn

[1] College of Forestry, Northwest A&F University, Yangling 712100, China

[2] Academy of Agriculture and Forestry Science, Qinghai University, Xining 810016, China

genes and the whole genome can be determined overall, and the regulatory mechanism in the genetic expression profile can be further revealed (Kumar et al. 2004; Shah and Gilchrist 2011; Zhou et al. 2014). Moreover, species with close genetic relationships will share a similar CUB, which would supply substantial evidence for identifying new germplasm resources and play a role in illustrating the evolutionary relationship among species (Ma et al. 2015). Furthermore, studies on CUB can be applied to predict the optimal heterologous expression receptor plant for target genes, as the transcription process (Zhou et al. 2016), translation efficiency of proteins (Frumkin et al. 2018), and RNA toxicity (Mittal et al. 2018) can be affected by CUB. It is of great significance for constructing expression vectors and investigations on unsuspected functional genes (Zelasko et al. 2013; Quax et al. 2015).

The CUB was also found to differ in the cp. genomes of different species (Liu and Xue 2005). To increase the adaptability of cp. genetic transformation and the expression level of target nucleic acids, it was reasonable to study the CUB in the whole cp. genome. Researchers have also studied the CUB in cp. genomes in a few species, such as *Calligonum mongolicum* (Duan et al. 2020), *Panicum* species (Li et al. 2021), *Cyperus alternifolius*, *Thalia dealbata*, and *Canna indica* (Deng et al. 2021a). *Elaeagnus* plants were used as hedge plants in urban areas for their decorative aspects (aromatic flowers and glistening leaves), dryness resistance, adaptability to various soil and water environments,

contamination prevention, and ability to attract insects and birds. However, in recent years, their fruits have been found to be nutrient-rich, and other parts have been recognized for their medicinal values (Patel 2015). Consequently, in addition to being edible, parts of the plant were utilized in traditional medicine as heat-clearing, muscle relaxing, analgesic, anti-inflammatory, astringent, and antifungal agents (Bendaikha et al. 2014; Saboonchian et al. 2014). Even though researchers have sequenced the complete cp. genomes of various *Elaeagnus* species (Choi et al. 2015; Wang et al. 2017; Liu et al. 2019; Lu et al. 2022), CUB studies on these cp. genomes have not yet been published.

In this work, the CUB in the cp. genomes of nine *Elaeagnus* species, such as *Elaeagnus angustifolia*, *Elaeagnus glabra*, *Elaeagnus henryi*, *Elaeagnus loureirii*, *Elaeagnus macrophylla*, *Elaeagnus mollis*, *Elaeagnus multiflora*, *Elaeagnus pungens*, and *Elaeagnus umbellata*, was identified (Fig. 1). The base composition and optimal codons of each species were compared. Based on previous studies, a correlation analysis, neutrality plot, effective number of codons (Nc) plot, parity rule 2 (PR2) plot, and t-distributed Stochastic Neighbor Embedding (tSNE) reducing dimension clustering were performed to investigate their CUB. Furthermore, to infer the history of their evolutionary relationship and predict the best heterologous expression receptor plant, the relative synonymous codon usage (RSCU) values of the nine species were also calculated and cluster analysis was

Fig. 1 Figures of the nine *Elaeagnus* species. **A** *Elaeagnus angustifolia*. **B** *Elaeagnus glabra*. **C** *Elaeagnus henryi*. **D** *Elaeagnus loureirii*. **E** *Elaeagnus macrophylla*. **F** *Elaeagnus mollis*. **G** *Elaeagnus multiflora*. **H** *Elaeagnus pungens*. **I** *Elaeagnus umbellata*
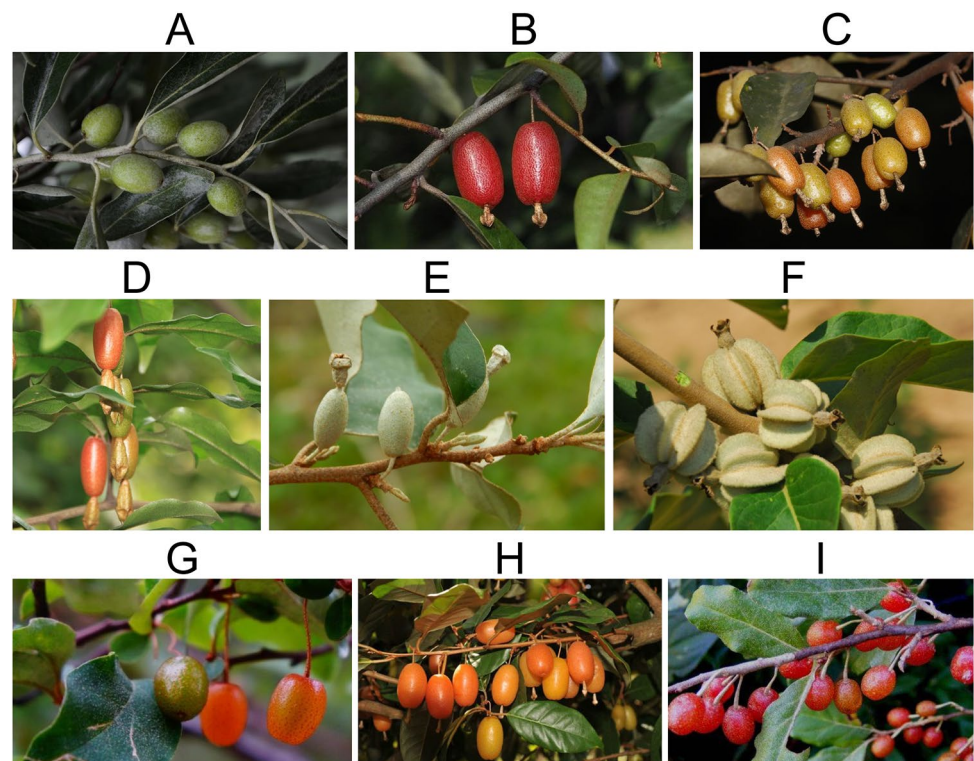
generated, which will provide theoretical support for the subsequent cp. genomic research of the *Elaeagnus* species.

Materials and methods

Flowchart

The flowchart of the materials and methods in this work can be seen in Fig. 2.

Sequence data

The entire cp. genomes of nine *Elaeagnus* species were downloaded from the GenBank database of the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/). Only the coding sequence (CDS), whose length was greater than or equal to 300 bp, was used for follow-up studies. The CDS was also required to be with ATG as its initiator codon and TAA, TAG, or TGA as its terminator codon. Additionally, the repeat sequences were discarded using the Perl language script (Table 1).

Statistical analysis of parameters related to CUB

Initially, the original data were obtained utilizing the codonW 1.4.2 software (http://codonw.sourceforge.net/). Moreover, equalizing values and homologous mobility scales of 15 parameters, such as aromaticity score (Aromo: the frequency of aromatic amino acids), codon adaptation index (CAI), codon bias index (CBI), general average hydropathicity (GRAVY), frequency of optimal codons (Fop), number of amino acids (L_aa), number of synonymous codons (L_sym), and effective number of codons (Nc), of the nine *Elaeagnus* species were calculated. Furthermore, the codons' GC, GC3s, GC12, T3s, C3s, A3s, and G3s (average base content) of each cp. genome were calculated via Excel 2019 and plotted into a bar chart on SPSS v26. The Pearson correlation analysis of the parameters above was also performed via the Sangerbox 3.0 cloud platform (Shen et al. 2022), taking the initial results of *E. mollis* as a representative.

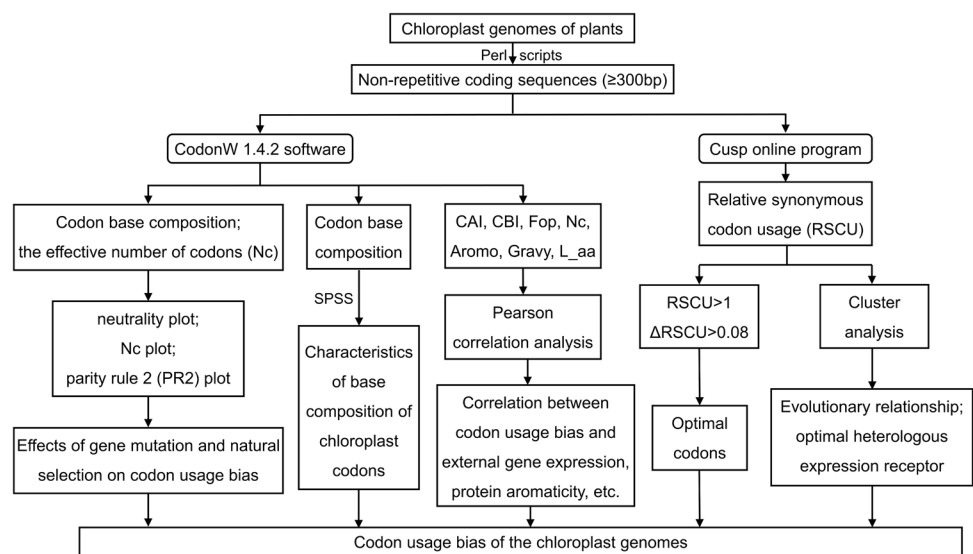**Fig. 2** The flowchart of the materials and methods used in this work

Chloroplast genomes of plants
Perl scripts
Non-repetitive coding sequences (≥300bp)

CodonW 1.4.2 software — Cusp online program

Codon base composition; the effective number of codons (Nc) | Codon base composition | CAI, CBI, Fop, Nc, Aromo, Gravy, L_aa | Relative synonymous codon usage (RSCU)

neutrality plot; Nc plot; parity rule 2 (PR2) plot

SPSS

Pearson correlation analysis

RSCU>1 ΔRSCU>0.08 | Cluster analysis

Characteristics of base composition of chloroplast codons

Correlation between codon usage bias and external gene expression, protein aromaticity, etc.

Optimal codons | Evolutionary relationship; optimal heterologous expression receptor

Effects of gene mutation and natural selection on codon usage bias

Codon usage bias of the chloroplast genomes

Table 1 Information of the chloroplast genomes of nine *Elaeagnus* species

No.	Genbank Accession	Species Latin Name	CDS (total)	CDS (filtered)
1	MK131395	*Elaeagnus angustifolia*	86	59
2	LC522137	*Elaeagnus glabra*	81	54
3	MZ846204	*Elaeagnus henryi*	83	58
4	MH394425	*Elaeagnus loureirii*	84	58
5	LC523635	*Elaeagnus macrophylla*	80	53
6	KY511611	*Elaeagnus mollis*	85	58
7	LC522136	*Elaeagnus multiflora*	79	53
8	MW145133	*Elaeagnus pungens*	85	59
9	LC522506	*Elaeagnus umbellata*	80	54

Optimal codons analysis

In general, the synonymous codons with more occurrences among the 64 codons were known as optimal codons (Ikemura 1985). The theoretical range of the Nc value is between 20 and 61 (included). Additionally, the bigger the Nc value, the broader the synonymous codon selective range, and the weaker CUB will be (and vice versa) (Wright 1990). The expectation value of the Nc was computed as:

$$ENc = 2 + GC3s + 29/[GC3s^2 + (1 - GC3s)^2].$$

The filtered CDSs were ordered by the Nc values at the beginning. Subsequently, 10% of the genes (approximately five CDSs for each species) were picked from both sides as high (with small Nc values) and low (with big Nc values) expression groups. Their homologous RSCU values were calculated utilizing the CUSP tool (https://www.bioinformatics.nl/cgi-bin/emboss/cusp) present in the European Molecular Biology Open Software Suite (EMBOSS) online software. The RSCU value of a codon was computed as:

$$RSCU = \frac{X_{ij}}{\sum_{j=1}^{n_i} X_{ij}} \cdot n_i$$

where X_{ij} stands for the probability that the jth codon appeared for the ith amino acid, and n_i is the number of codons encoding the ith amino acid (Chakraborty et al. 2020; Li et al. 2021). The theoretical range of RSCU is 0–6. Eventually, the codons that reached the condition (RSCU > 1 and ΔRSCU > 0.08) were regarded as the optimal codons (Table S1) (Zhang et al. 2007), and the results were represented by the UpSet plot using the Sangerbox 3.0 cloud platform.

Analysis of the CUB influencing factors

The dominant factor influencing the CUB in the cp. genes was estimated through three graph analysis methods: the neutrality plot, Nc plot, and PR2 plot, on SPSS v26 and Adobe Illustrator CC 2018 (AI 2018). The neutrality plot is the one-variable linear regression of GC12 on GC3s. In contrast, the Nc plot is the comparative study of actual Nc values with the specification curve of expected Nc values. Moreover, the PR2 plot illuminates the imbalance of A/T and G/C base mutations at the 3rd position of codons.

Evolutionary relationship analysis

Clustering analysis of the nine *Elaeagnus* species was initially performed based on the RSCU values via the complete linkage method and Euclidean distance using Sangerbox 3.0.

Furthermore, tSNE dimensionality reduction clustering analysis on the Sangerbox 3.0 platform was utilized for the nine *Elaeagnus* species sorted into Sect. *Deciduae* and Sect. *Sempervirentes* by the physiological and ecological characteristics of plant leaves adapting to the environment [first proposed by Servettaz (1909) and accepted by Chang (1983)]. In addition, the phylogenetic tree was constructed adopting the maximum likelihood estimation method (ML) via the FastTree v2.1.11 software (http://www.microbesonline.org/fasttree/), based on the filtered CDSs, in which sequence alignments were performed previously using the Multiple Alignment using Fast Fourier Transform (MAFFT) v7.480 procedure (https://mafft.cbrc.jp/alignment/software/windows.html).

Prediction of optimal heterologous expression receptor plant

Clustering analysis oriented to the typical *Elaeagnus* species (*E. mollis*) and model organisms, such as the whole genomes of *Bacillus subtilis*, *Escherichia coli*, *Microcystis aeruginosa*, *Saccharomyces cerevisiae*, and *Staphylococcus aureus*, together with the cp. genomes of *A. thaliana, Nicotiana sylvestris*, *T. aestivum*, and *O. sativa*, was carried out based on RSCU values acquired from the codon usage database (http://www.kazusa.or.jp/codon/) or calculated as previously done on the Sangerbox 3.0 cloud platform (Table S2).

Results and discussion

Statistical analysis of parameters related to the CUB

According to the mean values and corresponding variation ranges of the parameters closely associated with CUB, the major cp. CDSs were prone to mutation in the *Elaeagnus* species. Meanwhile, all the *rps7* genes of the nine cp. genomes were identical in sequence (Table S3). The cp. genomes of the nine *Elaeagnus* species possessed duplicated genes, such as *ndhB*, *rpl2*, *rps7*, *ycf1*, and *ycf2*, and specific genes not existing in all the nine sequence files or genes that were shorter than 300 bp, such as *clpP*, *ndhD*, *ndhE*, *petB*, *rpl16*, and *rps12*, which was parallel to that of former studies in the cp. genomes of the *Panicum* species (Li et al. 2021), C. *alternifolius*, *T. dealbata*, and *C. indica* (Deng et al. 2021a). It was also found that the longer the length of the single-copy gene, such as *ndhF*, *rpoC1*, *rpoB*, *psaA*, and *psaB*, the more stable the gene would be according to the variation ranges of Nc, which was also similar to previous research on the *Panicum* species (Table S3) (Li et al. 2021). It may indicate that the long single-copy genes

played an important role in the plant cells and were stable in sequences.

Most of the 20 amino acids are encoded by four codons whose 1st and 2nd bases are uniform. Therefore, CUB is reflected in the base composition at the 3rd position of codons in many cases (Zhao et al. 2019). The codons of the CDSs in the cp. genomes of the nine *Elaeagnus* species were found to favor ending with an A/T base. The mean G/C base contents of the 1st and 2nd positions were notably higher than that of the 3rd position, parallel to former studies on the CUB in the cp. genomes of plants, such as the *Lespedeza* species (Somaratne et al. 2019), *Camellia* species (Yengkhom et al. 2019)d *Mongolicum* (Duan et al. 2020). Nevertheless, the G3s values of the cp. genes in the *Elaeagnus* species appear to be higher than the species above, resulting in higher GC3s and GC values as well (Fig. 3A). It may lead to the higher stability of DNA structures in the cp. genomes of *Elaeagnus* species, as there are three hydrogen bonds between G and C bases.

In line with the output result of the Pearson correlation analysis, the CAI had a significantly positive correlation with the CBI ($p < 0.0001$, $r = 0.74$) and Fop ($p < 0.0001$, $r = 0.77$), indicating the association between CUB and external gene expression (Fig. 3B). The more optimal codons were used, the higher the gene expression level would be. Analogical results were also observed in the cp. genome of *Hemiptelea davidii* (Liu et al. 2020). Moreover, the CAI was positively correlated with T3s ($p < 0.0001$, $r = 0.52$) and C3s ($p < 0.01$, $r = 0.34$) and negatively correlated with A3s ($p < 0.001$, $r = -0.44$) and G3s ($p < 0.05$, $r = -0.31$), testifying that the cp.

genes finishing with a T/C base in *E. mollis* generally have a higher expression level. There was a significantly positive correlation between the CBI and C3s ($p < 0.001$, $r = 0.44$) and negative correlations between the CBI and G3s/(G3s + C3s) ($p < 0.0001$, $r = -0.58$), and G3s ($p < 0.001$, $r = -0.43$), indicating that the codons of cp. genes in *E. mollis* prefer C-termination rather than G-termination. In addition, the Nc had a significantly positive correlation with GC3s ($p < 0.001$, $r = 0.47$), which was also seen in parallel studies in the cp. genomes of *Porphyra umbilicalis* (Li et al. 2019) and *Mesona chinensis* (Tang et al. 2021) as well, and G3s ($p < 0.001$, $r = 0.47$), indicating that the more codons ending with G/C bases were applied, the weaker the CUB will be in the cp. genome of *E. mollis*.

The GRAVY had a significantly negative correlation with A3s/(A3s + T3s) ($p < 0.01$, $r = -0.41$) and A3s ($p < 0.05$, $r = -0.32$) and a significantly positive correlation with T3s ($p < 0.05$, $r = 0.37$), demonstrating that there were usually more T-terminated codons in the genes, from which the hydrophobicity of the translated proteins was relatively strong in the cp. genome of *E. mollis*. The Aromo had a significantly positive correlation with T3s ($p < 0.0001$, $r = 0.53$) and a significantly negative correlation with A3s/(A3s + T3s) ($p < 0.001$, $r = -0.46$) and GC ($p < 0.01$, $r = -0.41$), which indicates that TTT codons were more frequently used for phenylalanine and TAT codons were more frequently used for tyrosine in the cp. genes of *E. mollis*. Additionally, former studies have demonstrated that Axis 1 (the leading factor) acquired via correspondence analysis (COA) on the codon usage pattern of photosynthesis-associated genes
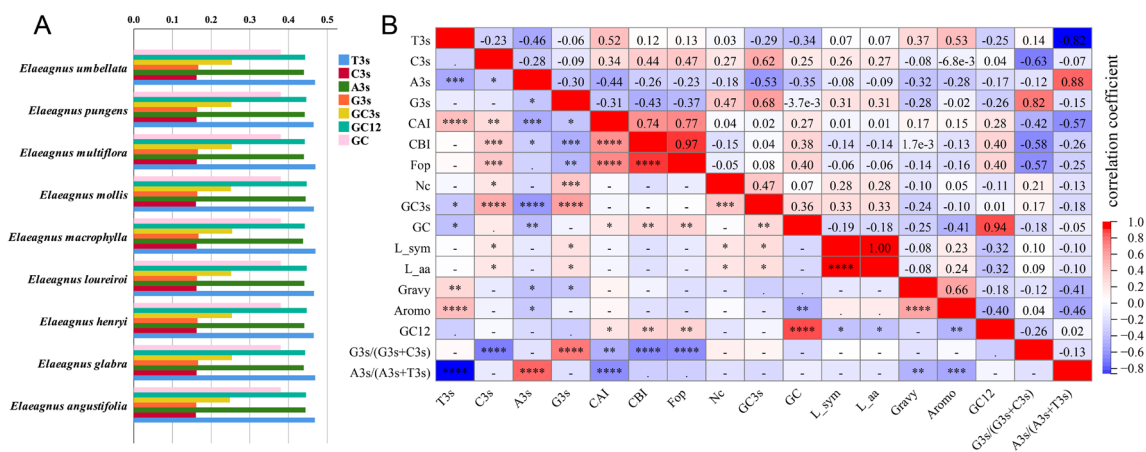

Fig. 3 Codon usage pattern of the nine *Elaeagnus* species. **A** Codon base composition of the nine *Elaeagnus* species. A3s/T3s/C3s/G3s, the frequency that codons have an A/T/C/G at their 3rd position of synonymous codons; GC3s, the G/C contents of the 3rd position of codons; GC12, the average G/C contents of the 1st and 2nd positions of codons; GC, the average G/C contents of the three positions of codons. **B** Pearson correlation analysis of parameters related to CUB towards the cp. genome of *Elaeagnus mollis*. CAI, codon adaptation index; CBI, codon bias index; Fop, frequency of optimal codons; Nc, effective number of codons; L_sym, number of synonymous codons; L_aa, number of amino acids; Gravy, general average hydropathicity; Aromo, aromaticity score (the frequency of aromatic amino acids); *Significant at $p < 0.05$ (two-tailed); **Significant at $p < 0.01$ (two-tailed); ***Significant at $p < 0.001$ (two-tailed); ****Significant at $p < 0.0001$ (two-tailed)

was observably correlated with the GRAVY, Aromo, and length of the deoxynucleotide chain but did not correlate with that of genetic system-related genes, supporting that the photosynthesis-associated genes play a pivotal role in the cp. genomes (Zhang et al. 2018).

Optimal codons analysis

Since the RSCU values of the codons in the cp. genes of *E. mollis* were analogous to the previous study on cp. genomes, this gives great credibility to the RSCU values calculated in this study (Cheng et al. 2020). Derived from the UpSet graph, the numbers of the optimal codons for the cp. genomes of the nine *Elaeagnus* species were no less than 15 and no more than 19 (Fig. 4). ATT, GAA, CGA, GTT, AAA, GTA, AGT, CGT, TTA, and GGT (10) were the optimal codons that synchronously arose in the nine *Elaeagnus* species, which signifies that these species all prefer using them. Additionally, the optimal codons for the cp. genomes of the nine species were comparatively similar, albeit with minor differences. Previous studies have suggested that the optimal codons for the cp. genome of *Populus alba* were CGT, GTC, TCT, and TTA, while those for *A. thaliana* were AAC, CGT, GGT, GTT, TAC, and TCA. The cp. genes of *T. aestivum* preferred to use AAA, ACT, CCT, CGT, GAG, GGT, TAC, and TCT, while the cp. genes of *Cycas taitungensis* favored AAT, CAT, CCA, GGT, GTA, TAT, TCA, TTA, and TTT as their codons (Zhou et al. 2008). Compared with the species mentioned above, the number of optimal codons for the cp. genomes of *Zea mays* (10), *Pinus koraiensis* (12) (Zhou et al. 2008), *Gynostemma* species (8–12) (Zhang et al. 2021), and the Euphorbiaceae species (17–18) (Wang et al. 2020) were more similar to that of the *Elaeagnus* species

(14–19). The large numbers of optimal codons may reflect their stronger CUB.

Analysis of the CUB influencing factors

Many factors influence CUB, such as the length of gene sequences (Marais and Duret 2001; Stoletzki 2011; Ribeiro et al. 2012), the codon position, protein translation efficiency (Haupt et al. 2009; Li et al. 2017), tRNA abundance (Buchan et al. 2006), gene mutations, and natural selection (Nandy 2002; Suzuki 2010). Nevertheless, researchers have discovered that base mutation and natural selection were the leading factors affecting CUB among different species (Fedorov et al. 2002; Hiraoka et al. 2009). In this work, *E. mollis* was used as an example to reveal the dominant factors influencing the CUB in the *Elaeagnus* species.

If GC12 was significantly correlated with GC3s in the neutrality plot, this would suggest that there was no difference in codon base usage between the 1st, 2nd, and the 3rd positions of codons, and mutation pressure was the leading factor influencing CUB. In contrast, if the correlation between GC12 and GC3s was not that significant, and the slope of the fitting curve was approaching zero, this would suggest that it is discrepant in the usage of the 1st, 2nd bases, and the 3rd codon bases, and the CUB was significantly correlated with the 3rd bases, which was strongly influenced by natural selection (Sueoka 1988; Liu and Xue 2004). Following the neutrality plot based on *E. mollis*, the correlation between GC12 and GC3s was insignificant ($R^2 = 1.764 \times 10^{-4}$). The slope ($k = 0.02$) was confoundedly approaching zero, demonstrating that it significantly differed in the usage of the 1st, 2nd, and the 3rd bases of its cp. genic codons. Natural selection strongly affected the

Fig. 4 Optimal codons of the nine *Elaeagnus* species. The UpSet graph shows the optimal codon sets of every species. The left displays the total optimal codon numbers of each species on the right. If there was a solid black dot in the middle part, it means that the species on the right has the codons above as its optimal codons
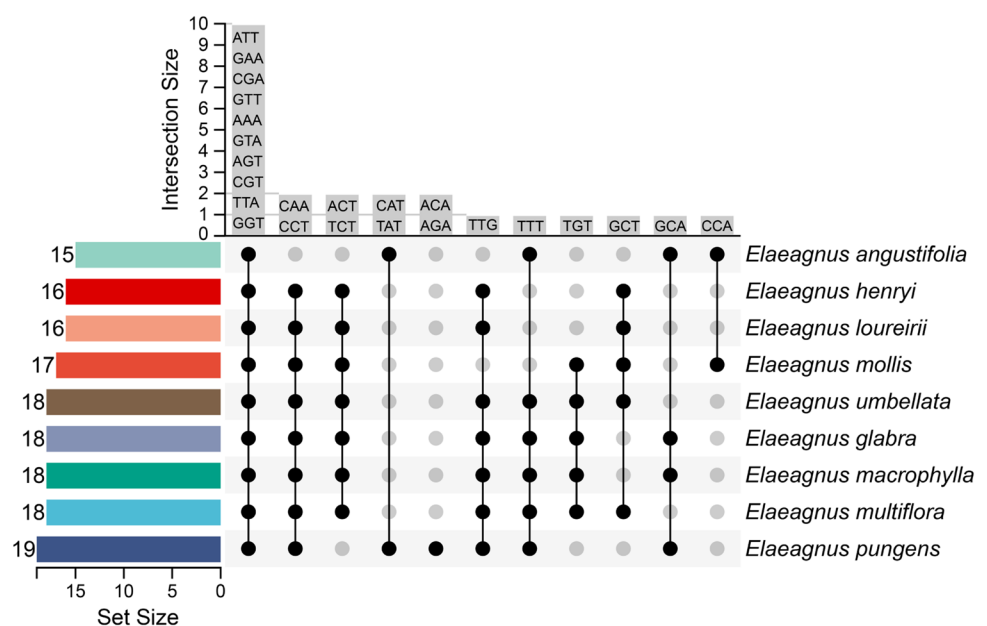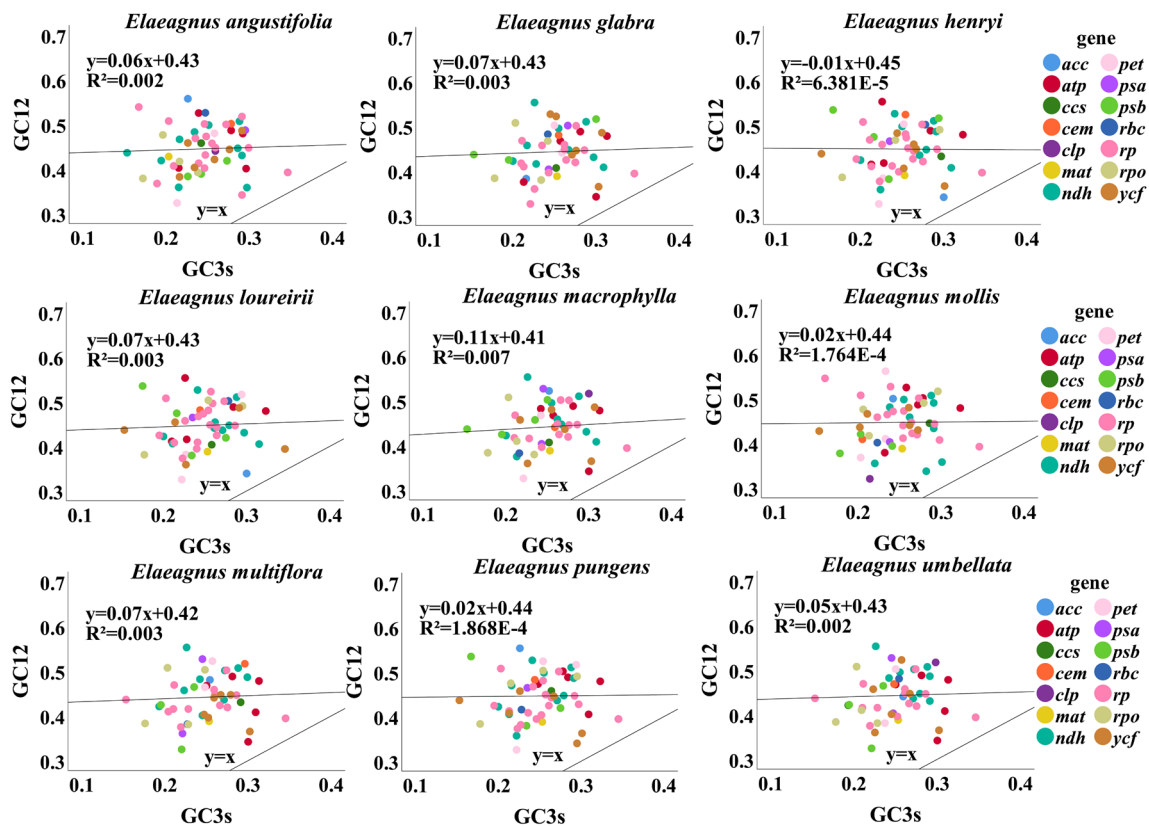
Fig. 5 Neutrality plot of the cp. genes (≥ 300 bp) in the nine *Elaeagnus* species. In the neutrality plot, a significant correlation between GC12 and GC3s suggests that there was no difference in codon usage between the 1st, 2nd, and the 3rd positions of codons and mutation pressure was the leading factor influencing codon usage bias. If on the contrary, this would suggest that it is discrepant in the usage of the 1st, 2nd bases, and the 3rd bases, and the codon usage bias was significantly correlated with the 3rd bases, which was strongly influenced by natural selection

CUB in the cp. genome (Fig. 5). Compared with the cp. genes of *Guizotia abyssinica* ($R^2 = 0.0282$, $k = 0.225$) and *Helianthus annuus* ($R^2 = 0.0293$, $k = 0.2388$) (Nie et al. 2014), the correlation coefficients and slopes of the neutrality plots based on the *Elaeagnus* species were evidently smaller (Fig. 5), suggesting that their CUB was more influenced by selection pressure than the two species above, with a strong preference. Different from the above result, the R^2 values and slopes of *Elaeagnus* species were closer to that of *Punica granatum* ($R^2 = 0.0036$, $k = 0.1165$) (Yan et al. 2019), *Triticum aestivum* ($R^2 = 0.0105$, $k = 0.1222$) (Zhang et al. 2007), *Ageratina adenophora* ($R^2 = 0.008$, $k = 0.1148$), and *Jacobeae vulgari* ($R^2 = 0.0057$, $k = 0.0809$) (Nie et al. 2014), indicating that their CUB strengths approach to one another. Among them, the GC12 of the cp. genes in *E. henryi* was negatively correlated with GC3s ($R^2 = 6.381 \times 10^{-5}$, $k = -0.01$), which is parallel to that of *Lactuca sativa* ($R^2 = 7 \times 10^{-6}$, $k = -0.0036$) (Nie et al. 2014), suggesting that the CUB in these two species were more than strongly influenced by natural selection. Their traits controlled by

cp. genes have been subjected to selection pressure for a period of time.

Provided that the data points representing genes were more abundantly located above the expected value curve in the Nc plot, their CUB was primarily influenced by the gene mutation. While the data points were more distributed far below the standard curve, selection pressure was the leading factor affecting the CUB in these genes (Wright 1990; Jia et al. 2009; Pan et al. 2009). According to the outcome of the Nc plot based on the cp. genomes of *E. mollis*, the CUB in *rpl2* (ribosomal protein gene) at the top of the graph was profoundly influenced by base mutation, with a weak preference (Fig. 6). Natural selection was the dominant factor affecting the CUB in the genes with a strong preference, such as *rps12* and *psbA* (PSII-A core protein of photosystem II), which were distributed far below the curve. In addition, the different cp. genes of *E. mollis* significantly differed in CUB. Compared with the *H. annuus* (Chen et al. 2021) and *Platycarya* species (Wang et al. 2021), the *Elaeagnus* species appeared to be more points above the curve in the Nc plot, demonstrating that the CUB in their cp. genomes
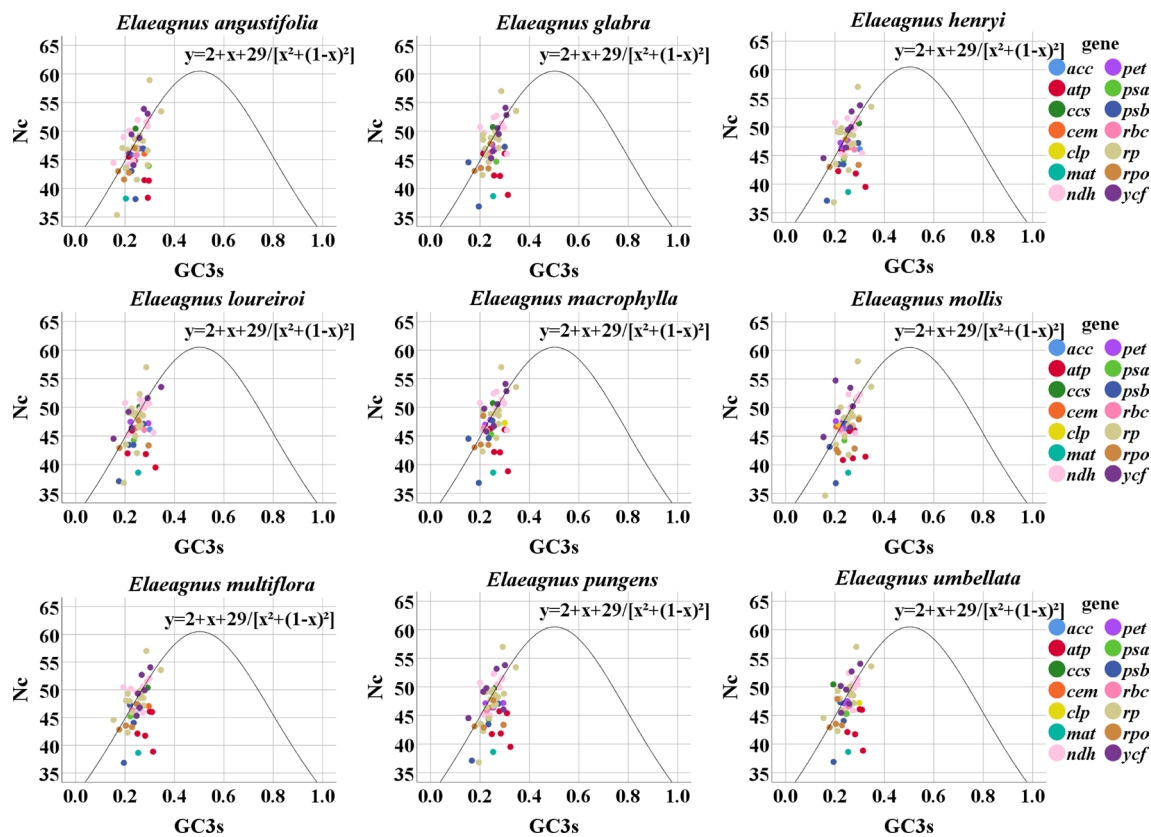
Fig. 6 Nc plot of the cp. genes (≥ 300 bp) in the nine *Elaeagnus* species. Provided that the data points representing genes were more abundantly located above the expected value curve in the Nc plot, the codon usage bias was primarily influenced by the gene mutation. While the data points were more distributed far below the standard curve, selection pressure was the leading factor affecting the codon usage bias of these genes

was considerably less influenced by selection pressure, with weaker preferences. Their cp. genes showed higher diversity, and the environment can accommodate this difference.

Similarly, the PR2 plot analysis was also one of the methods to determine the influence of mutation stress and natural selection on the CUB in genes (Sueoka 2001). By plotting a cruciform scatter graph with G3s/(G3s + C3s) as its abscissa and A3s/(A3s + T3s) as its ordinate, most of the points representing genes did not approach the center point, indicating that there were some other factors influencing the CUB in genes aside from genetic mutation, such as natural selection (Chakraborty et al. 2020; Tang et al. 2021). For genes close to the center point, such as *rbcL* (ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit) and *rpl22* among the cp. genes of *E. mollis*, their CUB was primarily influenced by the genetic mutation, with a weak preference. In the case of the points localizing the four corners of the plot, such as *rps12* (also found to be with a strong bias in the Nc plot method), *atpF*, *atpI* (ATP synthase), and *psaA* (PSI-A core protein of photosystem I) in the cp. genes of *E. mollis*, it verifies that other factors influence their CUB in addition to base mutation, such as selection pressure, and the preference

was strong (Fig. 7). Likewise, few genes closely approached to the central point in all *Elaeagnus* species in the PR2 plots, testifying that natural selection pressure played an essential role in the formation of their cp. gene CUB. The PR2 plot outcomes of the *Elaeagnus* species were similar to that of six Euphorbiaceae species (Wang et al. 2020). In comparison, the points representing the cp. genes of the species above were comparatively more distributed farther from the center point than that of *Coffea arabica* in the PR2 plots (Nair et al. 2012), revealing that the CUB in the cp. genomes of the *Elaeagnus* and Euphorbiaceae species were more affected by natural selection than *C. arabica*. Moreover, it suggested that their codon usage had more strong preferences.

Evolutionary relationship analysis

The RSCU values of different codons in distinct species manifest their evolutionary relationship in a sense, as codons associate gene sequences with polypeptide sequences. Additionally, they serve as supporting information in improving the taxonomic study of the *Elaeagnus* species (Li et al. 2019). Furthermore, to some extent, a cp. is matroclinously
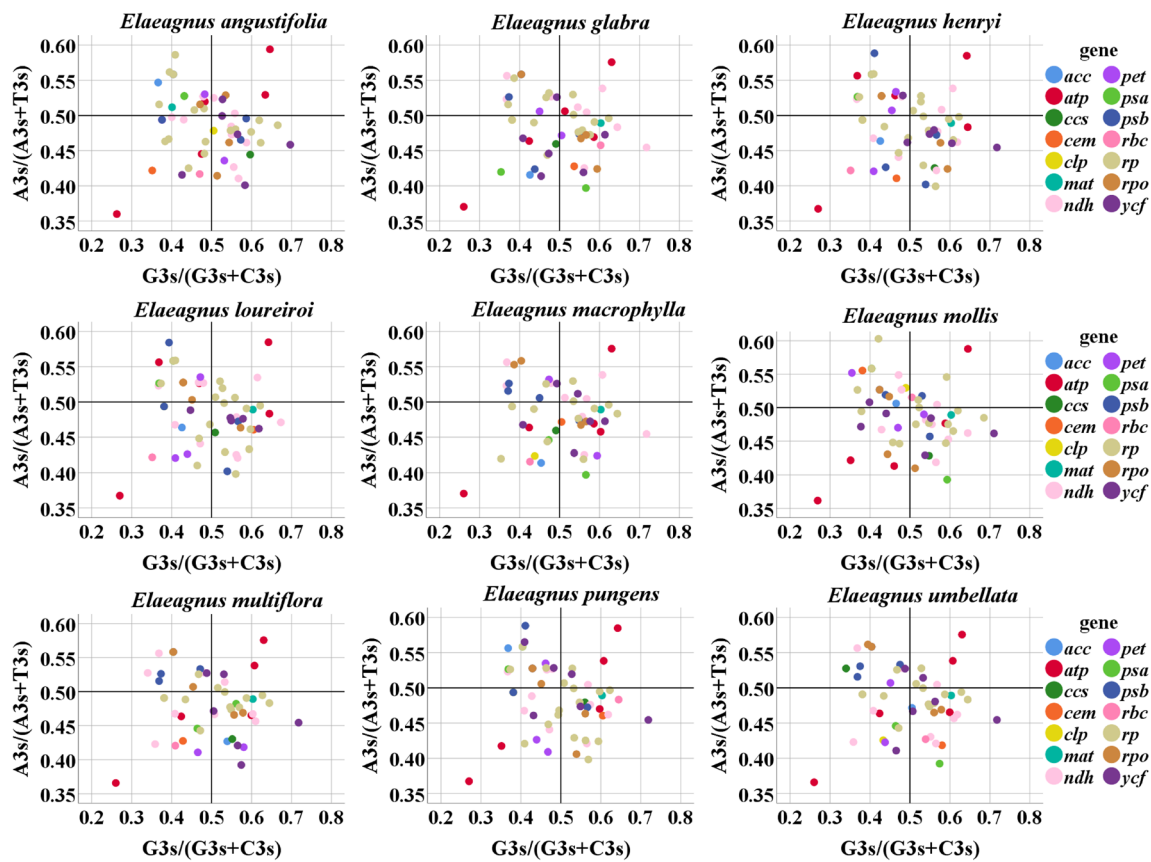
Fig. 7 PR2 plot of the cp. genes (≥ 300 bp) in the nine *Elaeagnus* species. Most of the points representing genes did not approach the center point, indicating that there were some other factors influencing the codon usage bias of genes aside from genetic mutation, such as natural selection

inherited, more suitable, and more convenient for phylogenetic analysis than the whole genome.

Following the clustering analysis based on the RSCU values, the nine *Elaeagnus* species were classified into five categories (Fig. 8A). Species in C0, C1, and C3 were all plants from Sect. *Deciduae*, while species in C2 and C4 were all from Sect. *Sempervirentes*. Moreover, the tSNE dimensionality reduction clustering analysis was also conducted in the nine *Elaeagnus* species grouped by evergreen and deciduous ecological characters based on their respective RSCU values. The trendlines of these two sections were whole separated (95% confidence interval), indicating that the RSCU values of the cp. genomes from these two sections differed from each other to a degree (Fig. 8B). In addition, it is possible for the following researchers to entirely distinguish the plants from these two groups based on the RSCU, which may help in the quick species identification of new genetic resources belonging to *Elaeagnus*.

The Sect. *Sempervirentes* species are evergreen erect or climbing shrubs with early flower opening and fruit ripening, while the Sect. *Deciduae* species are deciduous or semi-permanent green upright shrubs or trees with late

flower opening and fruit ripening (Servettaz 1909). It was also suggested that Sect. *Deciduae* should be renamed Sect. *Elaeagnus* (Sun and Lin 2010). In line with the former study, 15 *Elaeagnus* species were clustered into three branches dissimilar in the above biological and ecological traits based on the *matK* sequences, which were significantly stable in their cp. genomes, using ML analysis, adhering to the characteristics of the traditional taxonomic classification. In comparison, the *Elaeagnus* species could not be gathered smoothly for morphological clustering via principal component analysis (PCA). Additionally, the polygenetic trees based on *ITS* (nrDNA) sequences were also of significant discrepancy compared with that of *matK* genes via ML and maximum parsimony (MP) methods, and the *ITS* genes had duplex inheritance while the *matK* genes had matrilineal inheritance (Cheng et al. 2022). In this study, the identical outcome did not occur as well based on the whole cp. genomes via the same ML method, which may result from rapid changes in the CUB in some cp. genes in the *Elaeagnus* species affected by their local environment. Moreover, the fruit of *E. mollis* was significantly different from that of other species since it was the only one with eight ridges, and the species was
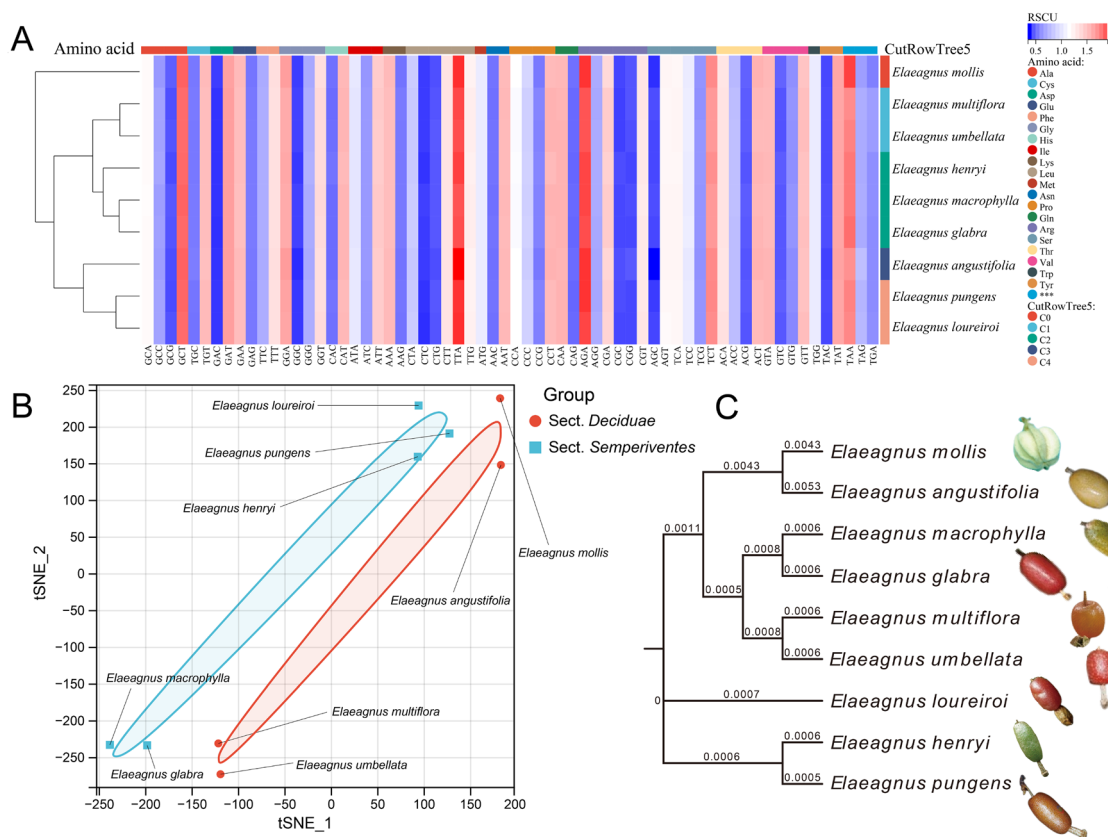
Fig. 8 Evolutionary relationship analysis of the nine *Elaeagnus* species. **A** The clustering analysis of the nine *Elaeagnus* species based on the RSCU values. **B** The tSNE dimensionality reduction clustering analysis of the nine *Elaeagnus* species classified by sections based on the RSCU values. **C** The phylogenetic tree of the nine *Elaeagnus* species based on CDS sequences with their fruits ($\geq$ 300 bp). The clustering analysis was performed via the complete linkage method and Euclidean distance. The closer the species are, the more similar their codon usage biases are

also found to be distant from other in the clustering analysis derived from the RSCU values (Fig. 8A, C).

Additionally, the CUB in the cp. genomes of the nine *Elaeagnus* species was extraordinarily similar to one another as the RSCU values of their codons were discovered to be tremendously close (Fig. 8A). Combining the clustering heatmap and the tSNE dimensionality reduction clustering based on the RSCU values with the phylogenetic tree generated from CDSs, it was found that the species clustered into one group in the clustering heatmap relatively approached one another in the other two clustering results as well, except for *E. henryi* (Fig. 8). In addition, the tSNE dimensionality reduction clustering was more similar to the ML-based phylogenetic tree, suggesting that the tSNE clustering method based on RSCU was more appropriate for evolutionary relationship analysis on the *Elaeagnus* species than the complete linkage method.

Prediction of optimal heterologous expression receptor plant

In genetic engineering research, such as external gene expression, molecular breeding, and functional verification, the matching degree of foreign genes and receptor genomes is essential for the successful acquirement of transgenic materials. Considerable distinctions between their CUB will probably create methylation hotspots, resulting in the silencing or diminished expression of external genes (Perlak et al. 1990). In addition to influencing the translation speed and folding of proteins, CUB can also affect transcriptional regulation at the mRNA level (Chen et al. 2017) and the expression of exogenous genes (Zhou et al. 2018). Based on the clustering heatmap, the RSCU values of the codons in the cp. genome of *E. mollis* were more adjacent to that of *A. thaliana*, indicating that a higher expression level would be attained by way of adopting the cp. genome of *A. thaliana* as the heterologous expression vector for the cp. genes of *E. mollis* (Fig. 9). In comparison, based on previous studies, the cp. genomes of *A. thaliana*, *Populus trichocarpa*,
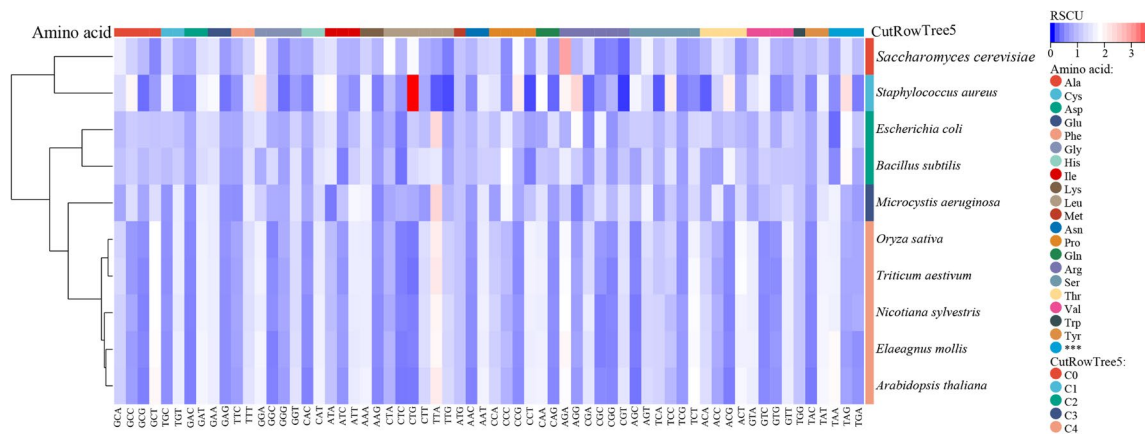
Fig. 9 The clustering analysis compared with model organisms based on the RSCU values. The clustering analysis was performed via the complete linkage method and Euclidean distance. The closer the species are, the more similar their codon usage biases are

and *S. cerevisiae* can be regarded as compatible heterogeneous expression receptor plants for the cp. genes of the *Miscanthus* species (Sheng et al. 2021) and Euphorbiaceae (Wang et al. 2020). Since the three species had similar optimal heterogeneous expression receptor plants, it may also suggest that they have a closer kinship in a sense. Astonishingly, the whole genome of *M. aeruginosa* (Cyanophyta) was clustered more closely to the cp. genomes of plants (C3 and C4), thus sustaining the endosymbiont hypothesis on the genesis of cp. to some extent.

Conclusion

Most of the cp. genes in the nine *Elaeagnus* species were prone to mutation, while the *rps7* gene sequences were synchronously selfsame. Selection pressure more significantly impacted CUB than gene mutation. Furthermore, the CUB in the cp. genes of the nine *Elaeagnus* species was extremely strong but with observable diversity among multifarious genes. The nine *Elaeagnus* species preferred using ten codons: ATT, GAA, CGA, GTT, AAA, GTA, AGT, CGT, TTA, and GGT. The *rps12* gene in the cp. genome of *E. mollis* had extraordinarily strong CUB via both the Nc and PR2 plot methods. Clustering outcomes based on the RSCU and cp. gene sequences were generally accordant, both of which could reveal the evolutionary relationship to a degree. In this work, it was suggested that the cp. genome of *A. thaliana* should be selected as the optimal heterologous expression receptor plant to obtain a higher expression efficiency in the following research on the cp. genes of the *Elaeagnus* species. Nevertheless, particular genes require further analysis due to the apparent distinctness of CUB in the different cp. genes of the *Elaeagnus* species.

References

Anwar AM, Aljabri M, El-Soda M (2021) Patterns of genome-wide codon usage bias in tobacco, tomato and potato. Biotechnol Biotechnol Equip 35:657–664. https://doi.org/10.1080/13102818.2021.1911684

Bendaikha S, Gadaut M, Harakat D, Magid A (2014) Acylated flavonol glycosides from the flower of *Elaeagnus angustifolia* L. Phytochemistry 103:129–136. https://doi.org/10.1016/j.phytochem.2014.03.025

Buchan JR, Aucott LS, Stansfield I (2006) tRNA properties help shape codon pair preferences in open reading frames. Nucleic Acids Res 34:1015–1027. https://doi.org/10.1093/nar/gkj488

Chakraborty S, Yengkhom S, Uddin A (2020) Analysis of codon usage bias of chloroplast genes in *Oryza* species. Planta 252:1–20. https://doi.org/10.1007/s00425-020-03470-7

Chang CY (1983) *Elaeagnus* L. In: Hong DY (ed) Flora reipublicae popularis sinicae, vol 52. Science Press, Beijing, pp 1–60

Chen SY, Li K, Cao WQ, Wang J, Zhao T, Huan Q, Yang YF, Wu SH, Qian WF (2017) Codon-resolution analysis reveals a direct and context-dependent impact of individual synonymous mutations on mRNA level. Mol Biol Evol 34:2944–2958. https://doi.org/10.1093/molbev/msx229

Chen SY, Zhang H, Wang X, Zhang YH, Ruan GH, Ma J (2021) Analysis of codon usage bias in the chloroplast genome of *Helianthus annuus* J-01. Journal 792:012009. https://doi.org/10.1088/1755-1315/792/1/012009

Cheng YQ, Yang YC, Fu XN, Liu L, Jiang ZM, Cai J (2020) Plastid genomes of *Elaeagnus mollis*: comparative and

phylogenetic analyses. J Genet 99:1–10. https://doi.org/10.1007/s12041-020-01243-5

Cheng CS, Fan SQ, Wang CJ, Ye LL, Wang ZP, Huang HW (2022) Phylogenetic analysis of *Elaeagnus* L. in China: a basis for genetic improvement of a berry crop. Front Plant Sci 13:899079. https://doi.org/10.3389/fpls.2022.899079

Chiapello H, Lisacek F, Caboche M, Hénaut A (1998) Codon usage and gene function are related in sequences of *Arabidopsis thaliana*. Gene 209:GC1–GC38. https://doi.org/10.1016/S0378-1119(97)00671-9

Choi KS, Son O, Park S (2015) The chloroplast genome of *Elaeagnus macrophylla* and *trnH* duplication event in Elaeagnaceae. PLoS ONE 10:e0138727. https://doi.org/10.1371/journal.pone.0138727

Deng N, Liu C, Tian Y, Song Q, Niu Y, Ma F (2021a) Complete chloroplast genome sequences and codon usage pattern among three wetland plants. Agron J 113:840–851. https://doi.org/10.1002/agj2.20499

Deng N, Liu CX, Song QA, Yao M, Yang R, Ma FF, Tian YX (2021b) Codon usage pattern analysis of *Gnetum luofuense* (Gnetales) based on transcriptome data. Agron J 113:696–705. https://doi.org/10.1002/agj2.20383

Duan HR, Zhang Q, Yang HS, Tian FP, Hu Y, Wang CM, Lu Y, Yuan HJ, Cui GX (2020) Complete chloroplast genome of *Calligonum Mongolicum*: genome organization, codon usage pattern, phylogenetic relationships, comparative structure and adaptive evolution analysis. Res Square. https://doi.org/10.21203/rs.3.rs-49271/v1

Fedorov A, Saxonov S, Gilbert W (2002) Regularities of context-dependent codon bias in eukaryotic genes. Nucleic Acids Res 30:1192–1197. https://doi.org/10.1093/nar/30.5.1192

Frumkin I, Lajoie MJ, Gregg CJ, Hornung G, Church GM, Pilpel Y (2018) Codon usage of highly expressed genes affects proteome-wide translation efficiency. Proc Natl Acad Sci 115:E4940–E4949. https://doi.org/10.1073/pnas.1719375115

Haupt S, Ziegler A, Cowan G, Torrance L (2009) Studies of the role and function of barley stripe mosaic virus encoded proteins in replication and movement using GFP fusions. In: Hicks BW (ed) Viral applications of Green fluorescent protein. Methods in Molecular Biology™, vol 515. Humana Press, Totowa, NJ, pp 287–297

Hiraoka Y, Kawamata K, Haraguchi T, Chikashige Y (2009) Codon usage bias is correlated with gene expression levels in the fission yeast *Schizosaccharomyces pombe*. Genes Cells 14:499–509. https://doi.org/10.1111/j.1365-2443.2009.01284.x

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13–34. https://doi.org/10.1093/oxfordjournals.molbev.a040335

Jia RY, Cheng AC, Wang MS, Xin HY, Guo YF, Zhu DK, Qi XF, Zhao LC, Ge H, Chen XY (2009) Analysis of synonymous codon usage in the *UL24* gene of duck enteritis virus. Virus Genes 38:96–103. https://doi.org/10.1007/s11262-008-0295-0

Krasovec M, Filatov DA (2022) Codon usage bias in phytoplankton. J Mar Sci Eng 10:168–178. https://doi.org/10.3390/jmse10020168

Kumar S, Tamura K, Nei M (2004) MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform 5:150–163. https://doi.org/10.1093/bib/5.2.150

Li J, Li HY, Zhi JK, Shen CZ, Yang XS, Xu JC (2017) Codon usage of expansin genes in *Populus trichocarpa*. Curr Bioinform 12:452–461. https://doi.org/10.2174/1574893611666161008195145

Li GL, Pan ZL, Gao SC, He YY, Xia QY, Jin Y, Yao HP (2019) Analysis of synonymous codon usage of chloroplast genome in *Porphyra umbilicalis*. Genes Genomics 41:1173–1181. https://doi.org/10.1007/s13258-019-00847-1

Li G, Zhang L, Xue P (2021) Codon usage pattern and genetic diversity in chloroplast genomes of *Panicum* species. Gene 802:145866. https://doi.org/10.1016/j.gene.2021.145866

Liu QP, Xue QZ (2004) Codon usage in the chloroplast genome of rice (*Oryza sativa* L. ssp. *japonica*). Acta Agronomica Sinica 30:1220–1224. https://doi.org/10.1300/J064v24n01_09

Liu QP, Xue QZ (2005) Comparative studies on codon usage pattern of chloroplasts and their host nuclear genes in four plant species. J Genet 84:55–62. https://doi.org/10.1007/BF02715890

Liu H, Wang M, Yue W, Xing G, Ge L, Nie X, Song W (2017) Analysis of codon usage in the chloroplast genome of Broomcorn millet (*Panicum miliaceum* L.). Plant Sci J 35:362–371. https://doi.org/10.11913/PSJ.2095-0837.2017.30362

Liu J, Gong LD, Qi L, Liu ZY, Niu YF, Shi C (2019) The complete chloroplast genome of *Elaeagnus conferta* Roxb (Elaeagnaceae). Mitochondrial DNA Part B 4:2035–2036. https://doi.org/10.1080/23802359.2019.1617074

Liu HB, Lu YZ, Lan BL, Xu JC (2020) Codon usage by chloroplast gene is bias in *Hemiptelea davidii*. J Genet 99:1–11. https://doi.org/10.1007/s12041-019-1167-1

Lu Y, Ma Q, Xu XL, Wang ZM, Andrie I, Savitskaya T, Yuzuak S (2022) Characterization of the complete chloroplast genome of *Elaeagnus pungens* (Elaeagnaceae) and phylogeny within Elaeagnaceae. Mitochondrial DNA Part B 7:1213–1215. https://doi.org/10.1080/23802359.2022.2090291

Marais G, Duret L (2001) Synonymous codon usage, accuracy of translation, and gene length in *Caenorhabditis elegans*. J Mol Evol 52:275–280. https://doi.org/10.1007/s002390010155

Ma QP, Li C, Wang J, Wang Y, Ding ZT (2015) Analysis of synonymous codon usage in *FAD7* genes from different plant species. Genet Mol Res 14:1414–1422. https://doi.org/10.4238/2015.February.13.20

Mittal P, Brindle J, Stephen J, Plotkin JB, Kudla G (2018) Codon usage influences fitness through RNA toxicity. Proc Natl Acad Sci 115:8639–8644. https://doi.org/10.1073/pnas.1810022115

Nair RR, Nandhini MB, Monalisha E, Murugan K, Sethuraman T, Nagarajan S, Rao NSP, Ganesh D (2012) Synonymous codon usage in chloroplast genome of *Coffea arabica*. Bioinformation 8:1096. https://doi.org/10.6026/97320630081096

Nandy A (2002) Investigations on evolutionary changes in base distributions in gene sequences. Internet Electron J Mol Des 1:545–558. https://doi.org/10.1.1.84.9810

Nie XJ, Deng PC, Feng KW, Liu PX, Du XH, You FM, Song WN (2014) Comparative analysis of codon usage patterns in chloroplast genomes of the Asteraceae family. Plant Mol Biol Rep 32:828–840. https://doi.org/10.1007/s11105-013-0691-z

Pan T, Li D, Luo MC, Tang FQ, Tien P, Pan ZS (2009) Analysis of synonymous codon usage in classical swine fever virus. Virus Genes 38:104–112. https://doi.org/10.1007/s11262-008-0296-z

Patel S (2015) Plant genus *Elaeagnus*: underutilized lycopene and linoleic acid reserve with permaculture potential. Fruits 70:191–199. https://doi.org/10.1051/fruits/2015014

Perlak FJ, Deaton RW, Armstrong TA, Fuchs RL, Sims SR, Greenplate JT, Fischhoff DA (1990) Insect resistant cotton plants. Nat Biotechnol 8:939–943. https://doi.org/10.1038/nbt1090-939

Quax TE, Claassens NJ, Söll D, Van Der Oost J (2015) Codon bias as a means to fine-tune gene expression. Mol Cell 59:149–161. https://doi.org/10.1016/j.molcel.2015.05.035

Ribeiro AS, Häkkinen A, Lloyd-Price J (2012) Effects of gene length on the dynamics of gene expression. Comput Biol Chem 41:1–9. https://doi.org/10.1016/j.compbiolchem.2012.10.002

Saboonchian F, Jamei R, Sarghein SH (2014) Phenolic and flavonoid content of *Elaeagnus angustifolia* L. (leaf and flower). Avicenna J Phytomed 4:231–238. https://doi.org/10.22038/ajp.2014.1975

Servettaz C (1909) Monographie des Eléagnacées. Beihefte Botanischen Centralblatt 25:413–417

Shah P, Gilchrist MA (2011) Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and

genetic drift. Proc Natl Acad Sci 108:10231–10236. https://doi.org/10.1073/pnas.1016719108

Sheng JJ, She X, Liu XY, Wang J, Hu ZL (2021) Comparative analysis of codon usage patterns in chloroplast genomes of five *Miscanthus* species and related species. PeerJ 9:e12173. https://doi.org/10.7717/peerj.12173

Shen WT, Song ZG, Zhong X, Huang M, Shen DT, Gao PP, Qian XQ, Wang MM, He XB, Wang T (2022) Sangerbox: a comprehensive, interaction-friendly clinical bioinformatics analysis platform. iMeta 1(e36):31–36. https://doi.org/10.1002/imt2.36

Somaratne Y, Guan DL, Wang WQ, Zhao L, Xu SQ (2019) The complete chloroplast genomes of two *Lespedeza* species: insights into codon usage bias, RNA editing sites, and phylogenetic relationships in Desmodieae (Fabaceae: Papilionoideae). Plants 9:51. https://doi.org/10.3390/plants9010051

Stoletzki N (2011) The surprising negative correlation of gene length and optimal codon use-disentangling translational selection from GC-biased gene conversion in yeast. BioMed Cent Evol Biol 11:1–12. https://doi.org/10.1186/1471-2148-11-93

Sueoka N (1988) Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci 85:2653–2657. https://doi.org/10.1073/pnas.85.8.2653

Sueoka N (2001) Near homogeneity of PR2-bias fingerprints in the human genome and their implications in phylogenetic analyses. J Mol Evol 53:469–476. https://doi.org/10.1007/s002390010237

Sun M, Lin Q (2010) A revision of *Elaeagnus* L. (Elaeagnaceae) in mainland China. J Syst Evol 48:356–390. https://doi.org/10.1111/j.1759-6831.2010.00085.x

Suzuki Y (2010) Statistical methods for detecting natural selection from genomic data. Genes Genet Syst 85:359–376. https://doi.org/10.1266/ggs.85.359

Tang DF, Wei F, Cai ZQ, Wei YY, Khan A, Miao JH, Wei KH (2021) Analysis of codon usage bias and evolution in the chloroplast genome of *Mesona chinensis* Benth. Dev Genes Evol 231:1–9. https://doi.org/10.1007/s00427-020-00670-9

Wang WC, Chen SY, Zhang XZ (2017) Characterization of the complete chloroplast genome of *Elaeagnus mollis*, a rare and endangered oil plant. Conserv Genet Resour 9:439–442. https://doi.org/10.1007/s12686-017-0706-4

Wang ZJ, Xu BB, Li B, Zhou QQ, Wang GY, Jiang XZ, Wang CC, Xu ZD (2020) Comparative analysis of codon usage patterns in chloroplast genomes of six Euphorbiaceae species. PeerJ 8:e8251. https://doi.org/10.7717/peerj.8251

Wang XS, Wang YQ, Li SH, Liu YL, Zhu B (2021) Analysis of codon usage bias in the *Platycarya* chloroplast genome. Tree Genet Mol Breed 11:1–11. https://doi.org/10.5376/tgmb.2021.11.0001

Wright F (1990) The 'effective number of codons' used in a gene. Gene 87:23–29. https://doi.org/10.1016/0378-1119(90)90491-9

Yan M, Zhao XQ, Zhou JQ, Huo Y, Ding Y, Yuan ZH (2019) The complete chloroplast genomes of *Punica granatum* and a comparison with other species in Lythraceae. Int J Mol Sci 20:2886. https://doi.org/10.3390/ijms20122886

Yengkhom S, Uddin A, Chakraborty S (2019) Deciphering codon usage patterns and evolutionary forces in chloroplast genes of *Camellia sinensis* var. *assamica* and *Camellia sinensis* var. *sinensis* in comparison to *Camellia pubicosta*. J Integr Agric 18:2771–2785. https://doi.org/10.1016/S2095-3119(19)62716-4

Zelasko S, Palaria A, Das A (2013) Optimizations to achieve high-level expression of cytochrome P450 proteins using *Escherichia coli* expression systems. Protein Expr Purif 92:77–87. https://doi.org/10.1016/j.pep.2013.07.017

Zhang WJ, Zhou J, Li ZF, Wang L, Gu X, Zhong Y (2007) Comparative analysis of codon usage patterns among mitochondrion, chloroplast and nuclear genes in *Triticum aestivum* L. J Integr Plant Biol 49:246–254. https://doi.org/10.1111/j.1744-7909.2007.00404.x

Zhang RZ, Zhang L, Wang W, Zhang Z, Du HH, Qu Z, Li XQ, Xiang H (2018) Differences in codon usage bias between photosynthesis-related genes and genetic system-related genes of chloroplast genomes in cultivated and wild solanum species. Int J Mol Sci 19:3142. https://doi.org/10.3390/ijms19103142

Zhang PP, Xu WB, Lu X, Wang L (2021) Analysis of codon usage bias of chloroplast genomes in *Gynostemma* species. Physiol Mol Biol Plants 27:2727–2737. https://doi.org/10.1007/s12298-021-01105-z

Zhao CL, Peng LY, Wang X, Chen JL, Wang L, Chen H, Lai ZX, Liu SC (2019) Codon preference and evolutionary analysis of *AtGAI* gene in *Amaranthus* species. J Chin Agric Univ 24:10–22. https://doi.org/10.11841/j.issn.1007-4333.2019.12.02

Zhou M, Long W, Li X (2008) Patterns of synonymous codon usage bias in chloroplast genomes of seed plants. For Stud China 10:235–242. https://doi.org/10.1007/s11632-008-0047-1

Zhou JH, Ding YZ, He Y, Chu YF, Zhao P, Ma LY, Wang XJ, Li XR, Liu YS (2014) The effect of multiple evolutionary selections on synonymous codon usage of genes in the *Mycoplasma bovis* genome. PLoS ONE 9:e108949. https://doi.org/10.1371/journal.pone.0108949

Zhou ZP, Dang YK, Zhou M, Li L, Yu CH, Fu JJ, Chen S, Liu Y (2016) Codon usage is an important determinant of gene expression levels largely through its effects on transcription. Proc Natl Acad Sci 113:E6117–E6125. https://doi.org/10.1073/pnas.1606724113

Zhou ZP, Dang YK, Zhou M, Yuan HY, Liu Y (2018) Codon usage biases co-evolve with transcription termination machinery to suppress premature cleavage and polyadenylation. Elife 7:e33569. https://doi.org/10.7554/eLife.33569