RESEARCH ARTICLE Legume Research- An International Journal, Volume 48 Issue 5: 793-803 (May 2025)

Transcriptome-based *Medicago varia* Codon Preference Analysis

Liu Bai[1], Fengling Shi[1], Yingtong Mu[1], Yuanyuan Cui[1]

10.18805/LRF-827

ABSTRACT

Background: The study explored codon usage bias in *medicago varia* transcriptome coding sequences, aiming to provide data support for understanding gene expression and enhancing molecular breeding methods in *Medicago varia*.

Methods: In this study, *Medicago varia* was used as the research material and 11722 complete open reading frame sequences were screened from the transcriptome sequencing results. Codon usage patterns and preferences were analyzed using software such as CodonW, R and Excel.

Result: According to the 11722 complete open reading frame sequences screened from the transcriptome sequencing results, Codon was used to calculate the Codon preference index. A total of 41354958 Unigenes were obtained. A total of 80757 Unigenes (39.66%) were annotated, with the majority being annotated in the NR database. GO annotations include 20 Biological processes, 16 Cellular components and 11 molecular function GO terms, respectively. 25713 genes were annotated into KOG, including 25 subclasses. The content of average GC was 42.87%, respectively. The study revealed that the effective number of codons (ENC) ranged from 23.9 to 61.0. Mutation and selection affected the codon preference. *Medicago varia* preferred 30 codons and preferred the third base of the codon as A/U, so it can be inferred that Medicago prefers the third codon with A/U.

Key words: Codon usage bias, *Medicago varia*, Optimal codon, Transcriptome.

INTRODUCTION

Transcriptome sequencing is a technique used to transcribe all RNA molecules from a particular cell, tissue, or organism at a given time comprehensively and rapidly. This method has gained popularity in the fields of genetic breeding, pathological analysis and physiological mechanism research (Feng *et al.,* 2022). All functional states of a cell produce specific RNA, including mRNA and non-coding RNA, which are transcribed from that cell (Wang *et al.,* 2021). Genetic codons serve as the connection between the nucleic acid and an organism's protein to transmit information. The arrangement of triple codons results in a total of 61 codons in the organism, with 3 stop codons removed. These codons encode 20 amino acids, resulting in multiple codons corresponding to a single amino acid. These codons encode the same amino acid and are synonymous codons (Xi *et al.,* 2022). Intriguingly, synonymous codons are not randomly or equally used during encoding (Ma *et al.,* 2015). Over time, due to gene mutations and selection pressure, certain codons are used more frequently than others, resulting in what's known as codon usage bias (CUB). This bias is characterized by the common occurrence of using synonymous codons of different frequencies (Qin *et al.,* 2022). CUB is widespread across species and differs from species as a code within the genetic code or the second genetic code (Xu *et al.,* 2021; Liu *et al.,* 2022). The analysis of CUB will aid in understanding molecular evolution, environmental adaptation, genomic features and gene functional requirements of different species. It will also help elucidate the superior agronomic performance of cultivated species (Tang *et al.,* 2000; Zhang *et al.,* 2018).

[1]College of Grassland Science, Key Laboratory of Grassland Resources of the Ministry of Education, Key Laboratory of Forage Cultivation, Processing and Higher Efficient Utilization of the Ministry of Agriculture and Rural Affairs Inner Mongolia Agricultural University, Hohhot 010019, Inner Mongolian, P.R. China.

Corresponding Author: Fengling Shi, College of Grassland Science, Key Laboratory of Grassland Resources of the Ministry of Education, Key Laboratory of Forage Cultivation, Processing and Higher Efficient Utilization of the Ministry of Agriculture and Rural Affairs Inner Mongolia Agricultural University, Hohhot 010019, Inner Mongolian, P.R. China. Email: shifengling@imau.edu.cn

How to cite this article: Bai, L., Shi, F., Mu, Y. and Cui, Y. (2025). Transcriptome-based *Medicago varia* Codon Preference Analysis. Legume Research. 48(5): 793-803. doi: 10.18805/LRF-827.

Submitted: 06-08-2024 **Accepted:** 07-02-2025 **Online:** 22-03-2025

At present, codon usage bias analysis based on transcriptome data has been studied in the use of codons ending in A or U in the protein expression process of *Ginkgo biloba* (He *et al.,* 2016) and *Medicago ruthenica* (Peng *et al.,* 2024), *Helianthus annuus* WRKY Transcription factors (Gao *et al.,* 2022), *Paeonia lactiflora* (Wu *et al.,* 2015) and *Morus cathayana* (Kong *et al.,* 2017) prefer to use codons ending in A or T. Species differences in the preferred use of codons are demonstrated. The production of CUB is related to many factors such as gene expression level, translational initiation effect, base composition of the gene, GC content, gene length and RNA structure (Brenner *et al.,* 2002). The most frequently used codon is called the optimal codon. Through the determination of the optimal codon, the expression vector of genetic engineering can be

designed in a targeted manner and the expression amount of the target gene can be increased (Liang *et al.,* 2019; Wang *et al* ., 2024). Analyzing the codon preference and optimal codon of plants can provide valuable information about gene sequence characteristics such as base content and codon bias. This information can help us gain a better understanding of the development of biomolecules, the utilization of resources and the optimization of gene expression (Alfalahi *et al.,* 2022). It lays a strong theoretical foundation for exploring these topics. In breeding research, the optimal expression of genes can be achieved through the re-optimization and synthesis of genes and codon optimization is the key step to improve gene expression, including replacing rare codons, optimizing RNA secondary structure, adjusting GC content, avoiding restriction enzyme cleavage sites and deleting complex structures (such as hairpins and repetitive structures) to improve gene expression efficiency, so as to speed up the breeding process. The *Escherichia coli* codon GBS surface protein Lrrc was optimized and its expression in *E. coli* was effectively improved. After codon optimization in plants, the codon SpCas9 could be stably expressed in plants(Chen *et al.,* 2018). The optimization of Cas9 codon based on gene editing technology significantly increased the expression of Cas9 protein in bananas and the gene editing efficiency was increased by 4 times after protoplast transformation (Shi *et al.,* 2018).

Medicago varia Martin. cv. Caoyuan No.1 is an excellent variety of alfalfa artificially cultivated, which is widely distributed in semi-arid regions such as Northeast China, North China and Northwest China. Due to its well-developed root system and strong stress resistance, it not only plays the role of maintaining water and soil, windbreak and sand fixation but also is an excellent variety for the establishment of mixed artificial grassland grazing utilization (Zhang *et al.,* 2024). Based on the important ecological use and feed value of alfalfa, to effectively utilize the resources of alfalfa and comprehensively develop its potential value, basic research on its genome and transcriptome has become an urgent problem to be solved.

In this study, 11722 coding sequences (CDS) obtained by transcriptome sequencing were used as the research object and the codon composition and usage parameters were statistically analyzed by CodonW and Excel software. The codon use preference characteristics of the transcriptome of Caoyuan No. 1 alfalfa were revealed. It provides data support for future genetic variation, genetic mapping, gene mapping, molecular breeding and resource development.

MATERIALS AND METHODS

Medicago varia Martin. cv. Caoyuan No.1 used in the experiment was collected from the forage experimental field of Inner Mongolia Agricultural University, at 111°71′ E and 40°81′N. Full, uniform and consistent tissues were selected and tested in 2020-2021. Before sequencing, the material was placed in a cryovial and treated with liquid nitrogen for more than 20 minutes (30 minutes is appropriate), which is not easy to degrade. Samples are stored at -80°C for total RNA extraction and cDNA synthesis. Total RNA was using an RNA extraction kit, reverse transcription synthesizes the first cDNA reaction strand were sequenced using the IlluminaHiSeq™ 4000 and 6188391094 high-quality sequences were obtained after splicing, removing linkers and low-quality sequences. A total of 80,757 Unigenes were assembled by De novo for screening and 11,722 high-quality transcribed gene sequences were finally obtained for codon analysis.

To obtain comprehensive gene function information, Unigenes were compared with NR, KOG, GO, Swissprot, KEGG, TrEMBL and other databases by NCBI Blast+v2.60 (Altschul *et al.,* 1997). The GO functional annotation information was obtained based on the protein annotation results of Swissprot and TrEMBL (Gao *et al.,* 2023). The transcripts were aligned to the database and TransDecoder v3.0.1 was used to predict the coding sequence (CDS) of *Medicago varia* Martin. cv. Caoyuan No.1.

The codon adaptation index (CAI), the effective number of codons (ENC) and relative synonymous codon usage (RSCU) of coding sequences of *Medicago varia* Martin. cv. Caoyuan No.1 was calculated by Codon W software. The total GC content of the statistical sequence and the GC content of the nucleotide at codon position 3 (GC3), the GC content of codon positions 1 and 2 are expressed by GC1, GC2 and the average value of GC1 and GC2 is recorded as GC12.

Relative synonymous codon usage indicates the relative use probability of encoding amino acids between specific synonymous codons. If the codon has no preference, the RSCU value is 1 and if the RSCU value of a synonymous codon is greater than 1 or less than 1, it indicates that the codon is used relatively more or less (Liu *et al.,* 2009), according to the results of RSCU, the high-frequency codon of *Medicago varia* Martin. cv. Caoyuan No.1 transcriptome was identified.

The effective codon number ENC is an important reference index to measure the codon use preference, its value is between 20~61, the closer to 61, the weaker the codon use preference, when the ENC value is 61, the gene sample is completely random and the closer to 20, the stronger the codon use preference. ENC plotting was related to gene use bias and ENC plotting used ENC values and GC3s as ordinate and abscissa as two-dimensional scatter plots and to detect the effect of base composition on codon bias (Wright *et al.,* 1990). The formula for calculating the standard curve is as follows:

$$ENC = 2 + GC3S + \frac{29}{GC3S2} + \frac{(1 - GC3S)}{2}$$

The GC12 of the transcriptome codon of *Medicago varia* Martin. cv. Caoyuan No.1 was used as the ordinate and GC3 was used as the abscissa to plot the scatter plot and the linear fitting regression analysis was performed.

To assess the correlation between GC3 and GC12. It is used to measure the degree to which natural selection pressure and mutation affect codon use preference.

The bias analysis (PR2-plot) was based on A3/(A3+T3) as the abscissa and G3/(G3+C3) as the ordinate to reveal the base composition of each gene, which was used to evaluate the relationship between purines and pyrimidines in the codons of each gene (Quax *et al.,* 2015).

Comparison of codon use preference between Sativa and other organisms in Medicago varia Martin. cv. Caoyuan No.1. The occurrence frequency of each codon in the Medicago varia was obtained from the CodonW software and compared with that of *Arabidopsis thaliana, Glycine max, Nicotiana tabacum, Saccharomyces, Escherichia coli* (Duret *et al.,* 1999).

RESULTS AND DISCUSSION

Transcriptome sequencing with assembly

A total of 41843596 Raw Reads were obtained by sequencing the cDNA library of *Medicago varia* Martin. cv. Caoyuan No.1. Removing low-quality reads (Q value less than 20 and length less than 35 nt) and connectors yielded a total of 41354958 clean reads. High-quality sequences (sequences with more than 20 base masses) account for 98.84%. GC content accounted for 42.74% of the total base. The results showed that the transcriptome data of *Medicago varia* from Caoyuan No.1 met the quality standards of transcriptome data analysis. A total of 80757 Unigenes were obtained by assembly and redundancy of de novo, with an average length of 655 bp and an N50 value of 1067 bp and the length of Unigenes (Table 1, Table 2), the number of Unigenes between 200~300 nt was the largest, with 31256 entries, accounting for 38.70%, 45354 entries, accounting for 56.16% of 300~2 000 nt and 4147 entries, accounting for 5.14%, which were larger than 2 000 nt (Table 3).

Functional annotation of *Medicago varia* Martin. cv. Caoyuan No.1 unigene

The Unigene of *Medicago varia* Martin.cv.Caoyuan No.1 was annotated into 4 databases (NR, SwissProt, KOG and KEGG library). From the annotation results, it can be concluded that 52753 (65.32%) of Unigene were successfully annotated. NR and SwissProt had the largest number of successful annotations, with 49353 (93.55%) and 35292 (66.90%) annotations, 25713 (48.74%) annotations in KOG and 18589 (35.24%) annotations in KEGG (Table 4).

GO and KOG functional annotation classification

In the GO functional annotation classification of *Medicago varia* Martin. cv. Caoyuan No. 1 (Fig 1), biological processes were annotated to 20 GO terms, cell composition was annotated to 16 GO terms and molecular functions were annotated to 12 GO terms. According to the KOG note (Fig 2), 25 metabolic pathways were injected into 25713 Unigene, with the highest number being universal functional prediction, followed by post-translational modifications. The least number of Unigene annotations are extracellular and nuclear structures.

GC content analysis and neutral plot analysis

Codons play a crucial role in translating gene sequences into proteins and analyzing their usage patterns is essential for studying protein translation efficiency and functions (Liu *et al.,* 2020). Due to the presence of codons, they typically do not impact the coding of amino acids, thus altering the structure and function of proteins. However, over a long period of evolution, some codons have been used more frequently than other codons in genes, *i.e.*, secrets preference for the use of yards. Codon preference use is an adaptive selection formed by species in the long-term process of natural selection and evolution, which is mainly influenced by genetic mutation pressure and natural selection pressure (Sueoka *et al.,* 1988).

GC content is an important indicator of an organism's genome base composition. The type of amino acids may be affected by GC1 and GC2, but not by the 3rd base mutation. When there is no selective pressure, mutations do not typically lead to differences in the base content at the three positions of the codon. However, the codons themselves do influence this content, so there is a strong correlation between GC content and codon usage bias. (Sharp and Li, 1986).

Codon W was used for codon use preference analysis for the 11722 selected Unigenes. The results showed that the average values of T3s, A3s, G3s and C3s in the transcriptome of *Medicago varia* Martin. cv. Caoyuan No. 1 were 45.21%, 34.55%, 23.81% and 22.27%, respectively and the average total GC content was 42.87%, with a fluctuation range of 27.10%~69.60%. T(T3s) and A(A3s) were the most common codons of A, T, G and C, followed by G3s and C3s, which were 45.21% and 34.55%, respectively, followed by G3s and C3s (23.81%) and 22.27%, respectively. The average codon adaptation index (CAI) was 0.215 and the fluctuation range was 0.097~0.883. The average codon preference index (CBI) was -0.038 and

Table 1: Data filtering and statistical results.

Category	Before filter reads number	After filter reads number	Proportion(%)	GC(%)
Sample	41843596	41354958	98.84%	42.74%

Table 2: Transcriptome data splicing results of *Medicago varia* martin. cv. caoyuan No.1.

Category	Genes number	GC(%)	N50	Max length	Min length	Average length
Sample	80757	39.66	1067	13121	201	655

the fluctuation range was -0.379~0.868. The average optimal codon usage frequency (Fop) was 0.398 and the fluctuation range was 0.192~0.926. The aromatic amino acid (Aromo) ratio was 8.29%, the average protein hydrophobic level (GRAVY) was -0.331, the average amino acid number (L-aa) was 345.916 and the average synonymous amino acid (L-sym) was 333.625 (Table 5) - the average ENC of the transcriptome of *Medicago varia* Martin. cv. Caoyuan No.1 was 49.659 and the fluctuation range was 24.77~61. The above analysis showed the codon preference of codon 3 in *Medicago varia* Martin. cv. Caoyuan No.1 was not high, but the GC3 content of different genes varied more than the total GC content.

The analysis of the neutral plot results showed that the gene samples were concentrated on both sides of the regression line of the neutral map and the changing trend of GC3s and GC12 was consistent, so there was no difference in the base composition of the three codon positions, indicating that the main reason for the influence of the codon preference of the codon of *Medicago varia* Martin. cv. Caoyuan No.1 was mutation pressure (Fig 3).

In this study, we analyzed the codon bias of Medicago varia Martin. cv. Caoyuan No.1 found that 27 preferred to use codons ending in base A/U and only 3 preferred to use codons ending in base G/C, indicating that Medicago varia Martin. cv. Caoyuan No.1 preferred to use codons ending in base A/U (Sueoka *et al.,* 1999).

Analysis of the effective codon number of genes in *Medicago varia* Martin. cv. Caoyuan No.1

The ENC plot was plotted with GC3s as the x-axis and ENC as the y-axis (Fig 4). Codon GC3s are distributed between 0.094~0.914. The ENC value is between 23.9~61 and the closer to 61 the value is, the weaker the bias is. The mean effective codon number (ENC) of the transcriptome of *Medicago varia* Martin. cv. Caoyuan No.1 was 47.98, with a maximum of 61 and a minimum of 23.9. This indicates that only a few sequences are codon-biased. The above analysis showed that the codon preference of *Medicago varia* Martin. cv. Caoyuan No.1 was not high overall, but the codon preference of different genes was inconsistent. From the ENC-GC3s plot, it can be seen that most of the genes of *Medicago varia* Martin. cv. Caoyuan No.1 was walked around the standard curve, while a few genes were scattered far away from the standard curve,

Table 3: The proportion of Unigene in each length interval of *Medicago varia* martin. cv. caoyuan No.1.

The length of the unigene (nt)	Number of unigene	Proportion (%)
< 300	31256	38.70%
300~2000	45354	56.16%
> 2000	4147	5.14%

Table 4: Statistics of Unigene annotation rate of *Medicago varia* martin. cv. caoyuan No.1.

Datebase	Total unigenes	NR	Swissprot	KOG	KEGG
The number of unigene	80757	49353	35292	25713	18589
Annotation proportion (%)	52753	93.55	66.9	48.74	35.24

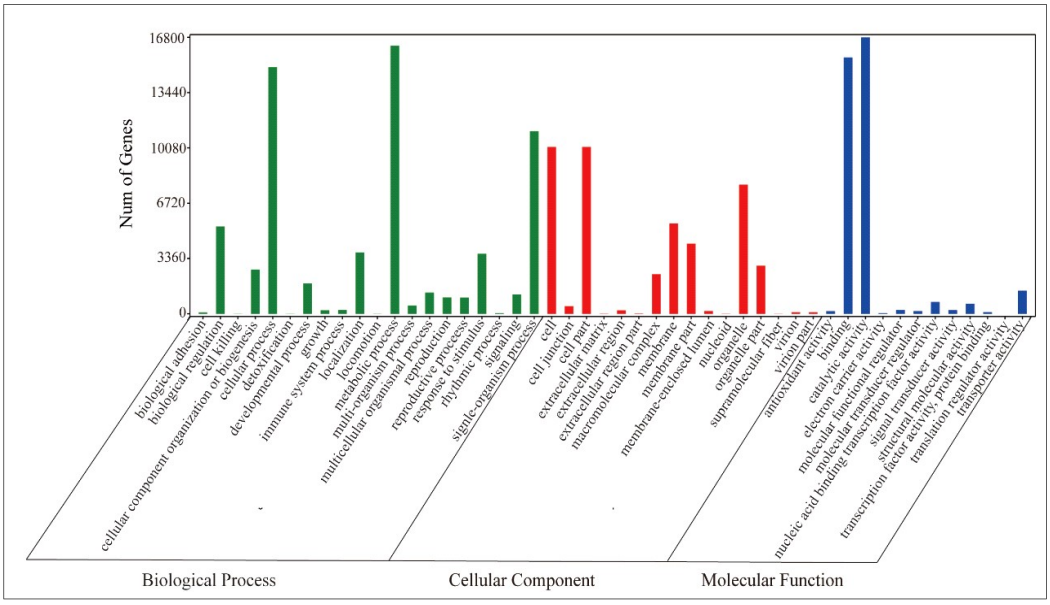


Fig 1: Go terms of *Medicago varia* Martin. cv. Caoyuan No.1.

indicating that mutational pressure, natural selection and some other factors led to the preference for the use of codons in *Medicago varia* Martin. cv.Caoyuan No.1.

The ENC value of the codon of *Medicago varia* Martin. cv. Caoyuan No.1 was significantly positively correlated with GC3s, indicating that the base composition affected the formation of codon bias of alfalfa gene to a certain extent and the ENC-plot results showed that the ENC value of most genes was close to the expected value.

PR2~plot bias analysis

The relationship between purines (A and G) and pyrimidines (T and C) at the third base of the codon of *Medicago varia* Martin. cv. Caoyuan No. 1 was analyzed by PR2~plot bias. The two straight lines in the PR2-plot plot divide the graph into four regions and the codons distributed in the upper half of the line indicate that the frequency of codon A is higher than that of T and vice versa, the frequency of use of T is higher than that of A. Codons in the left half of the line indicate that C is used more frequently than G and the right half of the line indicates that G is used more frequently than C. The frequency of use of T in *Medicago varia* Martin. cv. Caoyuan No.1 was higher than that of A and the frequency of G and C was relatively unequal (Fig 5) and the application frequency of the third base G of the codon of *Medicago varia* Martin. cv. Caoyuan No.1 was greater than that of C, indicating that the codon preference of *Medicago varia* Martin. cv. Caoyuan No.1 was affected by mutant pressure and natural selection (Oliveira *et al.,* 2021). Differences in species, ploidy and monodicots lead to differences in codon preference (Li *et al.,* 2024).

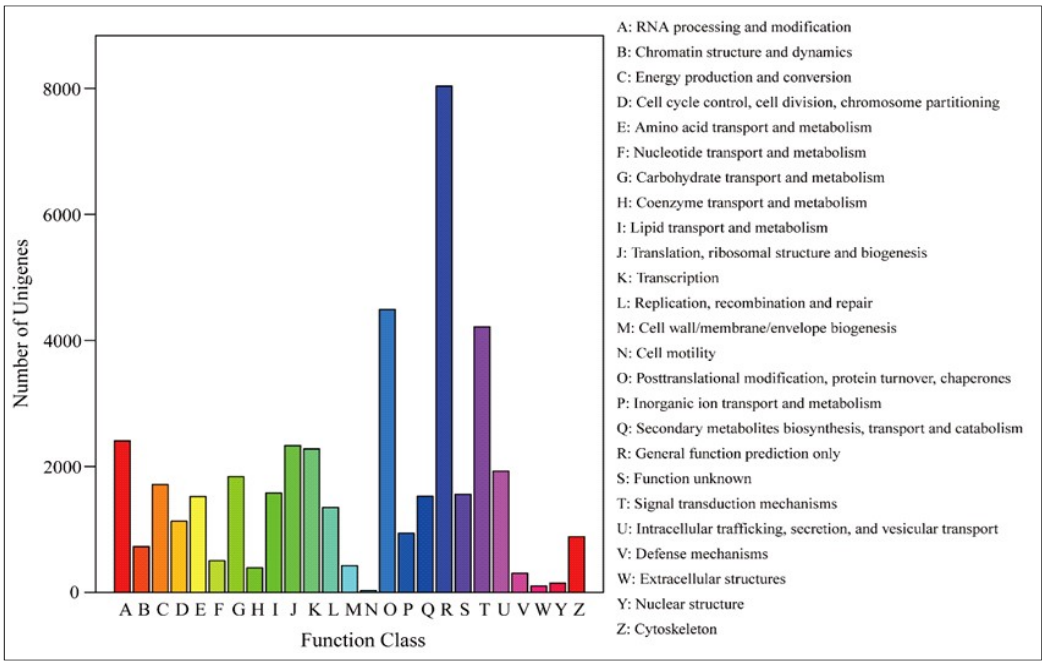


Fig 2: KOG function classification of *Medicago varia* martin. cv. caoyuan No.1.

Table 5: Codon base composition and parameters of *Medicago varia* martin. cv. caoyuan No.1 transcriptome.

Index	Average value	Maximum value	Minimum value
T3s	45.21%	73.42%	2.15%
C3s	22.27%	85.42%	2.35%
A3s	34.55%	64.08%	1.30%
G3s	23.81%	61.48%	0.00%
CAI	0.215	0.883	0.097
CBI	-0.038	0.868	-0.379
Fop	0.398	0.926	0.192
ENC	49.659	61	24.77
GC3s	35.75%	96.00%	10.20%
GC	42.87%	69.60%	27.10%
L_sym	333.625	1957	91
L_aa	345.916	2023	100
Gravy	-0.331	1.339604	-2.23826
Aromo	8.29%	24.59%	0.00%

Table 6: Relative usage statistics of the synonymous codon of *Medicago varia* Martin. cv. Caoyuan No.1.

Amino acid	Codon	Frequency of codon usage (1/1 000)						Analysis of codon bias				
		Mv	At	Gm	Nt	S	Ec	Mv/At	Mv/Gm	Mv/Nt	Mv/S	Mv/Ec
Ala	GCG	5.69	9.00	6.10	5.80	6.20	32.30	0.63	0.93	0.98	0.92	0.18*
	GCA	20.97	17.50	14.60	23.10	16.20	20.70	1.20	1.44	0.91	1.29	1.01
	GCT	29.34	28.30	49.70	31.20	21.20	15.40	1.04	0.59	0.94	1.38	1.91
	GCC	9.92	10.30	11.70	12.50	12.60	25.20	0.96	0.85	0.79	0.79	0.39*
Cys	TGT	10.62	10.50	5.90	9.80	8.10	5.20	1.01	1.80	1.08	1.31	2.04*
	TGC	6.33	7.20	11.70	7.20	4.80	6.40	0.88	0.54	0.88	1.32	0.99
Asp	GAT	39.98	36.70	32.20	36.90	37.60	32.80	1.09	1.24	1.08	1.06	1.22
	GAC	14.68	17.20	16.20	16.90	20.20	19.20	0.85	0.91	0.87	0.73	0.76
Glu	GAG	26.80	32.20	29.20	29.40	19.20	18.70	0.83	0.92	0.91	1.40	1.43
	GAA	38.29	34.30	32.20	36.00	45.60	39.30	1.12	1.19	1.06	0.84	0.97
Phe	TTT	26.01	21.80	40.90	25.10	26.10	22.20	1.19	0.64	1.04	1.00	1.17
	TTC	16.31	20.70	26.30	18.00	18.40	15.90	0.79	0.62	0.91	0.89	1.03
Gly	GGG	9.81	10.20	11.70	10.50	6.00	11.80	0.96	0.84	0.93	1.64	0.83
	GGA	21.58	24.20	23.40	23.20	10.90	8.90	0.89	0.92	0.93	1.98	2.42*
	GGT	23.95	22.20	14.70	22.30	23.90	24.20	1.08	1.63	1.07	1.00	0.99
	GGC	9.10	9.20	11.00	11.20	9.80	28.10	0.99	0.83	0.81	0.93	0.32*
His	CAT	16.06	13.80	14.40	13.40	13.60	12.80	1.16	1.12	1.20	1.18	1.26
	CAC	8.16	8.70	12.20	8.70	7.80	9.40	0.94	0.67	0.94	1.05	0.87
Ile	ATA	13.04	12.60	16.20	14.00	17.80	5.50	1.03	0.80	0.93	0.73	2.37*
	ATT	27.12	21.50	24.00	27.80	30.10	29.70	1.26	1.13	0.98	0.90	0.91
	ATC	13.01	18.50	11.60	13.90	17.20	23.90	0.70	1.12	0.94	0.76	0.54
Lys	AAG	30.94	32.70	40.90	33.50	30.80	11.00	0.95	0.76	0.92	1.00	2.81*
	AAA	33.71	30.80	33.80	32.60	41.90	34.00	1.09	1.00	1.03	0.80	0.99
Leu	TTG	23.44	20.90	25.30	22.30	27.20	13.00	1.12	0.93	1.05	0.86	1.80
	TTA	13.51	12.70	13.40	13.40	26.20	13.80	1.06	1.01	1.01	0.52	0.98
	CTG	9.35	9.80	10.40	9.80	10.50	51.10	0.95	0.90	0.95	0.89	0.18*
	CTA	8.71	9.90	8.20	9.40	13.40	3.90	0.88	1.06	0.93	0.65	2.23*
	CTT	25.60	24.10	18.80	24.00	12.30	11.40	1.06	1.36	1.07	2.08*	2.25*
	CTC	11.91	16.10	11.30	12.30	5.40	10.50	0.74	1.05	0.97	2.21*	1.13
Met	ATG	23.67	24.50	31.50	25.00	20.90	27.20	0.97	0.75	0.95	1.13	0.87
Asn	AAT	29.74	22.30	29.10	28.00	35.70	19.20	1.33	1.02	1.06	0.83	1.55
	AAC	17.95	20.90	17.90	17.90	24.80	21.70	0.86	1.00	1.00	0.72	0.83
Pro	CCG	5.63	6.90	3.60	5.00	5.30	22.40	0.82	1.56	1.13	1.06	0.25*

Table 6: Continue....

Table 6: Continue....

	CCA	17.92	16.20	19.20	19.80	18.30	8.40	1.11	0.93	0.91	0.98	2.13*
	CCT	20.44	18.70	9.00	18.70	13.50	7.20	1.09	2.27*	1.09	1.51	2.84*
	CCC	6.03	5.30	2.80	6.60	6.80	5.60	1.14	2.15*	0.91	0.89	1.08
Gln	CAG	14.72	15.20	18.60	15.00	12.10	29.40	0.97	0.79	0.98	1.22	0.50
	CAA	22.62	19.50	25.50	20.70	27.30	14.70	1.16	0.89	1.09	0.83	1.54
Arg	AGG	11.93	11.00	15.10	12.20	9.20	1.80	1.08	0.79	0.98	1.30	6.63*
	AGA	16.80	19.00	17.80	16.00	21.30	2.90	0.88	0.94	1.05	0.79	5.79*
	CGG	3.70	4.90	2.60	3.70	1.70	6.20	0.76	1.42	1.00	2.18*	0.60
	CGG	3.70	4.90	2.60	3.70	1.70	6.20	0.76	1.42	1.00	2.18*	0.60
	CGT	8.75	9.00	5.90	7.50	6.40	20.20	0.97	1.48	1.17	1.37	0.43*
	CGC	4.73	3.80	5.30	3.90	2.60	20.80	1.25	0.89	1.21	1.82	0.23*
Ser	AGT	15.55	14.00	14.10	13.30	14.20	9.40	1.11	1.10	1.17	1.09	1.65
	AGC	8.94	11.30	12.20	10.00	9.80	16.00	0.79	0.73	0.89	0.91	0.56
	TCG	5.78	9.30	3.60	5.30	8.60	8.80	0.62	1.61	1.09	0.67	0.66
	TCA	21.96	18.30	15.50	17.60	18.70	8.10	1.20	1.42	1.25	1.17	2.71*
	TCT	25.38	25.20	13.40	20.00	23.50	8.70	1.01	1.89	1.27	1.08	2.92*
	TCC	11.41	11.20	7.20	10.20	14.20	8.90	1.02	1.59	1.12	0.80	1.28
Thr	ACG	4.21	7.70	4.20	4.50	8.00	15.00	0.55	1.00	0.93	0.53	0.28*
	ACA	17.34	15.70	32.50	17.40	17.80	8.10	1.10	0.53	1.00	0.97	2.14*
	ACT	18.91	17.50	13.90	20.30	20.30	9.10	1.08	1.36	0.93	0.93	2.08*
	ACC	10.82	10.30	6.90	9.70	12.70	22.80	1.05	1.57	1.12	0.85	0.47*
Val	GTG	15.58	17.40	22.80	16.70	10.80	26.20	0.90	0.68	0.93	1.44	0.59
	GTA	10.17	9.90	9.70	11.40	11.80	10.90	1.03	1.05	0.89	0.86	0.93
	GTT	31.06	27.20	28.50	26.80	22.10	18.10	1.14	1.09	1.16	1.41	1.72
	GTC	9.10	12.80	9.00	11.10	11.80	14.80	0.71	1.01	0.82	0.77	0.62
Trp	TGG	11.86	12.50	14.20	12.20	10.40	15.30	0.95	0.84	0.97	1.14	0.78
Tyr	TAT	17.51	14.60	20.50	17.80	18.80	16.50	1.20	0.85	0.98	0.93	1.06
	TAC	10.26	13.70	13.50	13.50	14.80	12.30	0.75	0.76	0.76	0.69	0.83
Stop	TAA	0.90	0.90	0.40	1.10	0.80	1.10	1.00	2.25*	0.82	1.13	0.82
	TAG	1.20	0.50	0.60	0.50	0.70	0.50	2.40*	2.00*	2.40*	1.71	2.40*
	TGA	1.30	1.20	0.70	1.00	0.50	0.70	1.08	1.86	1.30	2.60*	1.86

Mv: *Medicago varia*; At: *Arabidopsis thaliana*; Gm: *Glycine max*; Nt: *Nicotiana tabacum L.*; S: *Saccharomyces*; Ec: *Escherichia coli*; *: The ratio was ≤0.5 or ≥2.0.

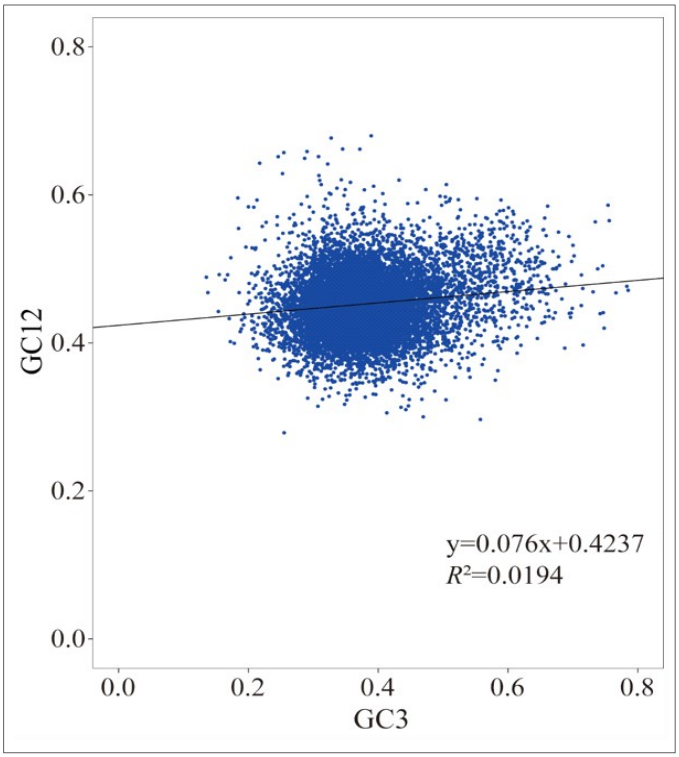


Fig 3: Analysis of transcriptome neutral mapping of *Medicago varia* Martin. cv. Caoyuan No.1.

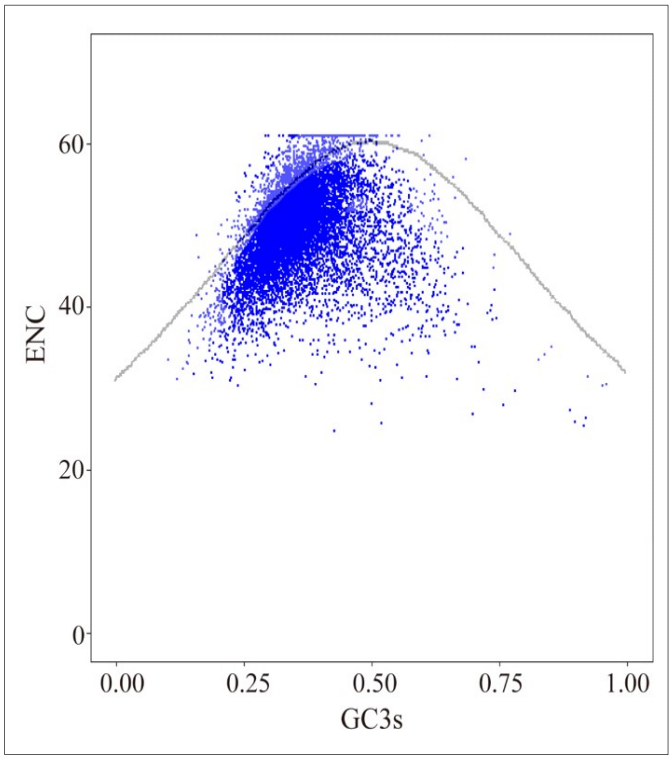


Fig 4: ENC-GC3s mapping analysis of *Medicago varia* Martin. cv. Caoyuan No.1.

Comparison of codon preference between alfalfa and representative species

If the ratio is 0.5~2, it indicates that the preference for the use of this codon is similar between the two species and *vice versa*. It was found that the codon preference of *Medicago varia* Martin. cv. Caoyuan No.1 was similar to that of the model species and there were 1, 4, 1, 5 and 24 codons with ratios outside the range of 0.5~2 (Table 6). The results indicated that there were different levels of differences between the codons of alfalfa and these model organisms, which were less different from Arabidopsis thaliana and tobacco and the most different from *Escherichia coli.*

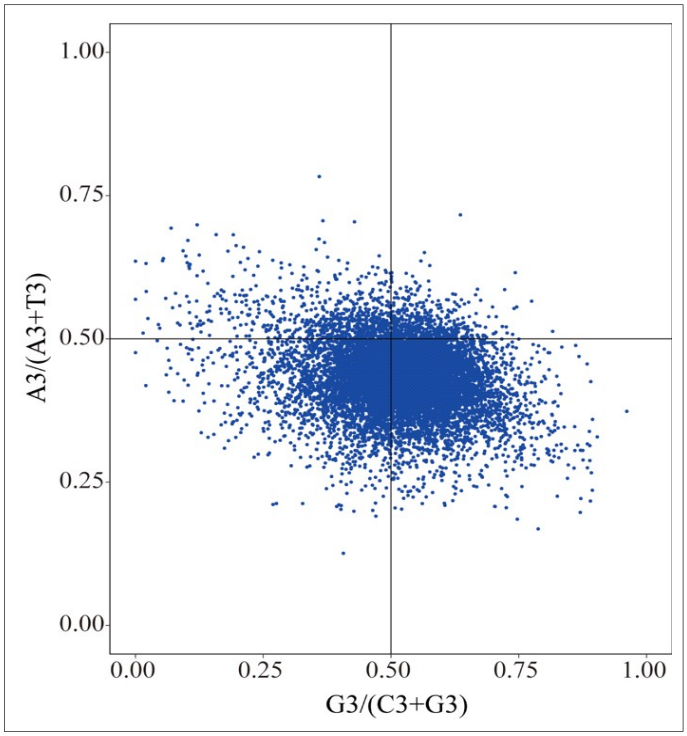


Fig 5: Analysis of PR2 bias plot.

	SECOND POSITION																			
	Base	Amino acid	Coden	High expression	Low expression	Amino acid	Coden	High expression	Low expression	Amino acid	Coden	High expression	Low expression	Amino acid	Coden	High expression	Low expression	Base	
		U				C				A				G					
FIRST POSITION	U	UUU*	Phe	1.39	0.94	UCU*	Ser	1.69	1.04	UAU*	Tyr	1.39	0.96	UGU*	Cys	1.48	0.93	U	THIRD POSITION
		UUC		0.61	1.06	UCC		0.37	1.45	UAC		0.61	1.04	UGC		0.52	1.07	C	
		UUA*	Leu	1.18	0.62	UCA*		2	0.83	UAA*	Ter	1.58	0.75	UGA	Ter	0.92	1.5	A	
		UUG*		1.76	0.75	UCG		0.13	0.82	UAG		0.5	0.75	UGG	Trp	1	1	G	
	C	CUU*		1.81	0.84	CCU*	Pro	1.76	0.82	CAU*	His	1.54	0.94	CGU	Arg	0.4	2.17	U	
		CUC		0.41	1.27	CCC		0.2	0.66	CAC		0.46	1.06	CGC		0.1	2.13	C	
		CUA*		0.66	0.27	CCA*		1.92	0.86	CAA*	Gln	1.6	0.77	CGA		0.23	0.5	A	
		CUG		0.18	2.24	CCG		0.12	1.66	CAG		0.4	1.23	CGG		0.11	0.55	G	
	A	AUU*	Ile	1.6	1.32	ACU*	Thr	1.64	0.73	AAU*	Asn	1.37	0.76	AGU*	Ser	1.3	0.66	U	
		AUC		0.5	1.47	ACC		0.44	1.91	AAC		0.63	1.24	AGC		0.51	1.2	C	
		AUA*		0.91	0.21	ACA*		1.84	0.54	AAA	Lys	1.11	1.31	AGA*	Arg	3.38	0.4	A	
		AUG	Met	1	1	ACG		0.09	0.82	AAG*		0.89	0.69	AGG*		1.79	0.25	G	
	G	GUU*	Val	2.12	1.22	GCU*	Ala	2.08	0.82	GAU*	Asp	1.63	1.15	GGU*	Gly	1.71	1.36	U	
		GUC		0.25	1.01	GCC		0.27	1.2	GAC		0.37	0.85	GGC		0.26	1.56	C	
		GUA*		0.67	0.5	GCA*		1.59	0.76	GAA	Glu	1.32	1.28	GGA*		1.54	0.54	A	
		GUG		0.96	1.27	GCG		0.07	1.22	GAG		0.68	0.72	GGG		0.49	0.54	G	

Fig 6: Optimal codon of the genome of *medicago varia* martin. cv. caoyuan No.1. *: RSCU value is greater than 1.

Comparing the codon preference between *Medicago varia* and Tobacco and *Escherichia coli*, it was found that *Medicago varia* and Tobacco were slightly different from each other, while *Escherichia coli* was quite different, indicating that *Arabidopsis thaliana* and Tobacco were suitable as recipients for genetic transformation in the verification of gene function of *Medicago varia* Martin. cv. Caoyuan No.1. Codon optimization can improve plant gene expression efficiency (Zhou *et al.,* 2016). Therefore, when the gene of *Medicago varia* Martin. cv. Caoyuan No.1 is exogenously expressed, it can be efficiently expressed in *Arabidopsis thaliana* and tobacco through codon optimization. Compared with *Escherichia coli,* there was little difference in codon preference between Saccharomyces cerevisiae and Saccharomyces cerevisiae codon, indicating that Saccharomyces cerevisiae was preferred as the protein expression system in Saccharomyces cerevisiae Medicago varia Martin. cv. Caoyuan No.1.

Optimal codon analysis of synonymous codons

This study compared the high and low codon gene expression sample libraries of *Medicago varia* Martin. cv. Caoyuan No.1 (Fig 6) and screened out 30 optimal codons, namely UUU, UUA, UUG, CUU, CUA, AUU, AUA, GUU, GUA, UCU, UCA, CCU, CCA, ACU, ACA, GCU, GCA, UAU, UAA, CAU, CAA, AAU, AAG, GAU, UGU, AGU, AGA, AGG, GGU, GGA. Among these 30 optimal codons, except for UUG and AAG, Outside of AGG, the third base of the remaining codon is all A/U, indicating that the *Medicago varia* Martin. cv. Caoyuan No.1 prefers the third base of the A/U codon, which is consistent with the above results.

Conclusion

Mutational pressure, natural selection and several other factors contribute to the preference for *Medicago varia* codon use. *Medicago varia* preferred 30 codons and preferred the third base of the codon as A/U, so it can be inferred that Medicago prefers the third codon with A/U.

Acknowledgement

The present study was supported by a Major demonstration project of “the open competition” for seed Industry science and technology innovation in Inner Mongolia (2022JBGS0016) and the Science and Technology Innovation 2030-Major Project (2022ZD04012). Fengling Shi designed the experiments, Liu Bai did the experiments and wrote the manuscript and others contributed equally to the manuscript, Yingtong Mu and Yuanyuan Cui carried out data statistics and drawing.

Disclaimers

The views and conclusions expressed in this article are solely those of the authors and do not necessarily represent the views of their affiliated institutions. The authors are responsible for the accuracy and completeness of the information provided, but do not accept any liability for any direct or indirect losses resulting from the use of this content.

Informed consent

All animal procedures for experiments were approved by the Committee of Experimental Animal care and handling techniques were approved by the University of Animal Care Committee.

Conflict of interest

The authors declare that there are no conflicts of interest regarding the publication of this article. No funding or sponsorship influenced the design of the study, data collection, analysis, decision to publish, or preparation of the manuscript.

References

Alfalahi, A.O., Hussein, Z.T., Khalofah, A., Sadder, M.T., Qasem, J.R., Al-Khayri, J.M., Jain, S.M., Almehemdi, A.F. (2022). Epigenetic variation as a new plant breeding tool: A review. Journal of King Saud University Science. 34: 102302.

Altschul, S., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller, W., Lipman D.J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Research. 25(17): 3389-3402.

Brenner, S., (H), J.M., Broughton, W.J. (2002). Encyclopedia of Genetics. Codon Usage Bias. pp: 402-406.

Chen, K. (2018). Optimization of CRISPR Technology and Modification of PAMRecognition Sites of Cas9 Enzyme. (Master), Chinese Academy of Agricultural Sciences.

Duret, L., Mouchiroud, D. (1999). Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila* and *Arabidopsis*. Proceedings of the National Academy of Sciences of the United States of America. USA. 96(8): 4482-4487.

Feng, C., Shan, M., Xia, Y., Zheng, Z., He, K., Wei, Y., Song, K., Meng, T., Liu, H., Hao, Y., Liang, Z., Wang, Y., Huang, Y. (2022). Single-cell RNA sequencing reveals distinct immunology profiles in human keloid. Frontiers in Immunology. 13: 940645.

Gao, Y., Lu, Y., Song, Y., Jing, L. (2022). Analysis of codon usage bias of WRKY transcription factors in *Helianthus annuus*. BMC Genomics Data. 23(1): 46.

Gao, Y., Zhang, D., Li, H. (2023). The Professional Go Annotation Dataset. IEEE Transactions on Games. 15(4): 517-526.

He, B., Dong, H., Jiang, C., Cao, F., Tao, S., Xu, L. (2016). Analysis of codon usage patterns in *Ginkgo biloba* reveals codon usage tendency from A/U-ending to G/C-ending. Scientific Reports. 6: 35927.

Kong, W.Q., Yang, J.H. (2017). The complete chloroplast genome sequence of *Morus cathayana* and *Morus multicaulis* and comparative analysis within genus *Morus* L. PeerJ. 5: e3037.

Li, L., Xu, M., Liu, Y. (2024). Codon Bias Analysis of Medicago Genome. Acta Agrestia Sinica. 23(09): 2695-2706.

Liang, E., Qi, M.J., Ding, Y.Q., Zhang, L. (2019). Analysis of codon bias in *Panax japonicus* transcriptome. Jiangsu Nongye Xuebao (Jiangsu Agricultural Sciences). 47(2): 67-71.

Liu, H., He, R., Zhang, H., Huang, Y., Tian, M., Zhang, J. (2009). Analysis of synonymous codon usage in *Zea mays*. Molecular Biology Reports. 37(2): 677-684.

Liu, Y. (2020). A code within the genetic code: Codon usage regulates co-translational protein folding. Cell Communication and Signaling. 18: 145.

Ma, Q., Li, C., Wang, J., Wang, Y., Ding, Z. (2015). Analysis of synonymous codon usage in FAD7 genes from different plant species. Genetics and Molecular Research. 14(1): 1414-1420.

Oliveira, J.L., Morales, A.C., Hurst, L.D., Urrutia, A.O., Thompson, C.R.L., Wolf, J.B. (2021). Inferring adaptive codon preference to understand sources of selection shaping codon usage bias. Molecular Biology and Evolution. 38(8): 3247-3266.

Peng, X., Mu, Y., Wu, F., Fu, N., Shi, F., Zhang, Y. (2024). Analysis of codon preferences in *Medicago ruthenica* based on transcriptome data. Legume Research. doi: 10.18805/LRF-776.

Qin, L., Ding, S., Wang, Z., Jiang, R., He, Z. (2022). Host plants shape the codon usage pattern of Turnip Mosaic Virus. Viruses. 14(10): 2267.

Quax, T.E.F., Claassens, N.J., Söll, D., van der Oost, J. (2015). Codon bias as a means to fine-tune gene expression. Molecular Cell. 59(2): 149-161.

Sharp, P.M., Li, W. (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. Journal of Molecular Evolution. 24(1/2): 28-38.

Shi H, Lin Y, Lai Z. (2018). Research progress on CRISPR/Cas9-mediated genome editing technique in plants. Chinese Journal of Applied and Environmental Biology. 24(3): 640-650.

Sueoka, N. (1988). Directional mutation pressure and neutral molecular evolution. Proceedings of the National Academy of Sciences of the United States of America. 85(8): 2653-2657.

Sueoka, N. (1999). Two aspects of DNA base composition: G+C content and Translation-Coupled Deviation from Intra-Strand Rule of A=T and G=C. Journal of Molecular Evolution. 49(1): 49-62.

Tang, L., Shah, S., Chung, L., Carney, J., Katz, L., Khosla, C., Julien, B. (2000). Cloning and heterologous expression of the epothilone gene cluster. Science. 287(5453): 640-642.

Wang, J., Li, X., Lu, Y., Huang, Q., Sun, Y., Cheng, M., Li, F., Shi, C., Zeng, Y., Wang, C., Cao, X. (2021). Analysis of lncRNAs and mRNA Expression in the ZBTB1 Knockout Monoclonal EL4 Cell Line and Combined Analysis With miRNAs and circRNAs. Frontiers in Cellular and Infection Microbiology. 11.

Wright, F. (1990). The 'effective number of codons' used in a gene. Gene. 87(1): 23-29.

Wang S, Shi F, Shi R, Zhang Y (2024). Seed dormancy and germination in alfalfa (*Medicago falcata* L.). Legume Research. 47(2): 234-241. doi: 10.18805/LRF-758.

Wu, Y., Zhao, D., Tao, J. (2015). Analysis of codon usage patterns in herbaceous peony (*Paeonia lactiflora* Pall.) based on transcriptome data. Gene. 6(4): 1125-1139.

Xi, Z., Davis, L., Baxter, K., Tynan, A., Goutou, A., Greiss, S. (2021). Using a quadruplet codon to expand the genetic code of an animal. Nucleic Acids Research. 50(9): 4801-4812.

Xu, B., Wu, R.N., Gao, C.P., Shi, F.L. (2021). Establishment of tissue culture regeneration system for *Medicago ruthenica* L. cv. 'Zhilixing'. Legume Research. 45(2): 162-167. doi: 10.18805/LRF-659.

Zhang, R., Zhang, L., Wang, W., Zhang, Z., Du, H., Qu, Z., Li, X., 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. International Journal of Molecular Sciences. 19(10): 3142.

Zhang, Y., Wu, X., Wang, X., Dai, M., Peng, Y. (2024). Crop root system architecture in drought response. Journal of Genetics and Genomics. 52(1): 4-13.

Zhou, Z., Dang, Y., Zhou, M., Li, L., Yu, C., Fu, J., Chen, S., Liu, Y. (2016). Codon usage is an important determinant of gene expression levels largely through its effects on transcription. Proceedings of the National Academy of Sciences of the United States sof America. p113.