RESEARCH ARTICLE Legume Research- An International Journal, Volume 49 Issue 3: 377-387 (March 2026)

Comparative Analysis of Codon Usage Bias in Nuclear Genome Coding Sequences (CDSs) Across 30 Angiosperm Species Centered on the Leguminosae (Fabaceae)

Jingshi Lu[1], Xinyue He[1], Yuehua Zhang[1], Fengling Shi[1]

10.18805/LRF-930

ABSTRACT

Background: Synonymous codon usage bias (CUB) modulates gene expression by influencing translational efficiency and accuracy, shaped by mutational biases and natural selection. A unified framework is essential to identify lineage-specific codon preferences and optimize expression and sequence design across species.

Methods: Nuclear CDSs from 30 angiosperm species, with a focus on Fabaceae, were analyzed using CodonW to compute gene-level indices (GC/GC3s, ENc, CAI, CBI, Fop). Codon preference patterns were examined through hierarchical clustering, PCA, and analyses of mutational and selective influences using ENc–GC3s plots, neutrality plots, and PR2-bias.

Results: Substantial variation in GC3s, ENc, and CAI/CBI/Fop was observed, with patterns aligned with phylogenetic differentiation. PCA revealed clear lineage-level clustering based on codon preferences, with Fabaceae species distinctly separated from Poaceae and other groups. Strong correlations between GC and GC3s, as well as between GC and CBI/Fop, highlighted the pivotal role of base composition in shaping codon preference. Additional analyses indicated that selective pressures, alongside mutational biases, contribute to codon usage, with expression-related indices providing evidence of translation-associated selection. This study establishes a comprehensive framework for understanding codon usage across 30 angiosperms, offering valuable insights for codon optimization and heterologous expression in the molecular breeding of legume crops and other species.

Key words: Codon usage bias, ENc, Fabaceae, Nuclear genome CDS, RSCU.

INTRODUCTION

Synonymous codon usage bias (CUB) is pervasive in coding sequences and represents a key phenotype linking molecular evolution to functional genomics. Although synonymous substitutions do not alter amino acid sequences, extensive evidence indicates that codon choice can modulate gene expression output and fitness by shaping translational elongation dynamics and decoding accuracy, co-translational folding, protein quality control and mRNA turnover (Plotkin and Kudla, 2011; Hanson and Coller, 2018; Presnyak *et al*., 2015). Accordingly, codon preference not only reflects genome composition and mutational spectra, but also provides an informative entry point for understanding gene regulation and adaptive evolution (Cheeran *et al.,* 2024; Basudha *et al.,* 2025; Zhang *et al.,* 2025).

Mechanistically, codon usage patterns are typically shaped by multiple forces, including mutational bias and base-composition constraints, natural selection associated with translational efficiency and accuracy (often coupled to tRNA availability and the translation apparatus) and processes such as strand asymmetry and recombination-associated biases (Sueoka, 1988; Hanson and Coller, 2018). These concepts have motivated a set of widely used quantitative indices and diagnostic frameworks: metrics such as CAI, ENc (Nc) and RSCU are applied to characterize bias strength and preferred codon spectra, whereas graphical diagnostics-including ENc-GC3s (ENC-plots), PR2-bias plots and neutrality plots-are commonly used to evaluate the relative contributions of mutational pressure and selection (Sharp and Li, 1987; Wright, 1990; Sueoka, 1992, 1995).

In angiosperms, pronounced lineage-level differences have been documented in GC3 architecture and synonymous codon-ending preferences. Notably, Poaceae often exhibit distinctive GC3 distributions accompanied by characteristic codon preference patterns, highlighting strong phylogenetic specificity in the evolution of codon usage (Kawabe and Miyashita, 2003; Tatarinova *et al*., 2010). Fabaceae (Leguminosae) is of major agricultural and ecological importance, encompassing diverse forage

[1]Key Laboratory of Grassland Resources/College of Grassland Science, Ministry of Education Inner Mongolia Agricultural University, Hohhot 010011, China.

Corresponding Author: Fengling Shi, Key Laboratory of Grassland Resources/College of Grassland Science, Ministry of Education Inner Mongolia Agricultural University, Hohhot 010011, China. Email: sfl0000@126.com

How to cite this article: Lu, J., He, X., Zhang, Y. and Shi, F. (2026). Comparative Analysis of Codon Usage Bias in Nuclear Genome Coding Sequences (CDSs) Across 30 Angiosperm Species Centered on the Leguminosae (Fabaceae). *Legume Research*. **49(3):** 377-387. doi: 10.18805/LRF-930.

Submitted: 10-01-2026 **Accepted:** 03-03-2026 **Online:** 11-03-2026

and grain/oil crops. In recent years, the rapid expansion and refinement of high-quality nuclear genome resources for representative legumes-such as soybean, *Medicago* spp., chickpea, diploid progenitors of peanut and Lotus japonicus-have provided a solid foundation for genome-wide comparative analyses (Schmutz *et al.*, 2010; Young *et al.*, 2011; Varshney *et al.*, 2013; Bertioli *et al.*, 2016; Sato *et al.*, 2008). Nevertheless, comprehensive comparisons of codon usage bias in nuclear CDSs that are centered on Fabaceae within a unified analytical framework and explicitly include key outgroups remain relatively limited. This gap constrains a systematic understanding of conserved regularities, lineage-specific differentiation axes and their potential mechanistic drivers.

To address this, we integrated nuclear genome CDSs from 30 plant species, centered on Fabaceae and supplemented with phylogenetically informative outgroups, to systematically delineate codon usage preference landscapes and their putative determinants across species. We computed gene-level indices including GC/GC3s, ENc (Nc), CAI/CBI/Fop and third-position nucleotide composition and constructed codon-level RSCU matrices to identify preferred codon sets. We further combined ENC-plots, PR2-bias analyses and neutrality plots to assess the relative contributions of mutational bias and selection and applied multivariate analyses to resolve shared versus lineage-specific codon preference patterns. This work provides a systematic evidence base for comparative codon usage evolution at the nuclear-genome scale in Fabaceae and offers transferable guidance for cross-species gene design, heterologous expression and codon optimization in molecular breeding.

MATERIALS AND METHODS

Study species and acquisition of CDS datasets

We performed a comparative analysis of codon usage bias in nuclear genome coding sequences (CDSs) across 30 angiosperm species, with Fabaceae (Leguminosae) as the focal lineage and Poaceae, Brassicaceae and additional phylogenetically relevant taxa included as outgroup references. Species abbreviations and Latin binomials are provided in Table 1. Nuclear CDS datasets were retrieved from publicly available resources, including NCBI, EnsemblPlants and Phytozome. To ensure data quality and cross-species consistency, we used annotated protein-coding gene sequences and applied a unified processing scheme to maximize comparability among species. All CDSs were converted to FASTA format for downstream analyses. The research was conducted at the Inner Mongolia Agricultural University in 2025.

Codon usage indices and computational analyses

For each species, codon-usage-related indices were calculated from the quality-filtered CDS dataset using CodonW (Peden, 1999). Specifically, we quantified: GC content (GC and GC3s), *i.e.,* the proportion of G and C nucleotides in the coding sequence, with GC3s representing GC content at third codon positions and serving as a key descriptor of codon-ending preferences (Wright, 1990); the effective number of codons (ENc, also denoted Nc), which measures the extent of synonymous codon diversity, with lower ENc values indicating stronger codon bias (Wright, 1990); CAI (Codon Adaptation Index) and CBI (Codon Bias Index), which evaluate, respectively, the extent to which a gene's codon usage conforms to a preferred/optimized codon set and the overall magnitude of codon bias (Sharp and Li, 1987) and Fop (Frequency of Optimal Codons), defined as the proportion of codons belonging to an optimal set and commonly used as a proxy for translation-associated optimization in highly expressed genes.

Relative synonymous codon usage (RSCU)

Relative synonymous codon usage (RSCU) was computed to quantify the extent to which a synonymous codon is used more or less frequently than expected under equal usage among synonymous alternatives. For a given amino acid, an RSCU value > 1 indicates preferential usage of the codon, whereas an RSCU value < 1 indicates avoidance

Table 1: Plant species included in this study.

Species	Latin	Clade
Atri	*Amborella trichopoda*	Basal_angiosperm
Ncol	*Nymphaea colorata*	Basal_angiosperm
Alyr	*Arabidopsis lyrata*	Brassicaceae
Atha	*Arabidopsis thaliana*	Brassicaceae
Esal	*Eutremasalsugineum*	Brassicaceae
Pper	*Prunus persica*	Eudicot
Ptri	*Populus trichocarpa*	Eudicot
Slyc	*Solanum lycopersicum*	Eudicot
Vvin	*Vitis vinifera*	Eudicot
Adur	*Arachis duranensis*	Fabaceae
Cari	*Cicer arietinum*	Fabaceae
Ccaj	*Cajanus cajan*	Fabaceae
Gmax	*Glycine max*	Fabaceae
Lang	*Lupinus angustifolius*	Fabaceae
Ljap	*Lotus japonicus*	Fabaceae
Mru	*Medicago ruthenica*	Fabaceae
Msat	*Medicago sativa*	Fabaceae
Mtru	*Medicago truncatula*	Fabaceae
Psat	*Pisum sativum*	Fabaceae
Pvul	*Phaseolus vulgaris*	Fabaceae
Tpra	*Trifolium pratense*	Fabaceae
Vrad	*Vigna radiata*	Fabaceae
Macu	*Musa acuminata*	Monocot
Bdis	*Brachypodium distachyon*	Poaceae
Hvul	*Hordeum vulgare*	Poaceae
Osat	*Oryza sativa*	Poaceae
Sbic	*Sorghum bicolor*	Poaceae
Sita	*Setaria italica*	Poaceae
Svir	*Setaria viridis*	Poaceae
Zmay	*Zea mays*	Poaceae

(Grantham *et al.*, 1980). We calculated RSCU for each codon in each species and performed cross-species comparisons using the resulting species-by-codon matrix.

ENc-GC3s relationship (ENC-plot)

The ENc-GC3s plot (ENC-plot) was used to assess whether codon bias can be largely explained by mutational bias in base composition. Under a mutation-driven model, ENc is expected to follow a characteristic theoretical relationship with GC3s (Wright, 1990). For each species, we plotted gene-wise (ENc, GC3s) values and compared the observed distribution against the theoretical expectation to infer the relative contribution of mutational pressure versus additional forces (*e.g.*, selection).

PR2-bias analysis (PR2-plot)

Parity Rule 2 (PR2) bias analysis was conducted to characterize deviations from parity in nucleotide usage at third codon positions, focusing on the balance between A vs T and G vs C. For each gene, we calculated A3/(A3 + T3) and G3/(G3 + C3) and visualized their distributions in PR2 plots to evaluate the extent to which mutational bias and/or selection influences codon usage patterns (Sueoka, 1995).

Neutrality plot analysis

Neutrality plot analysis was performed by examining the relationship between GC12 (mean GC content at the first and second codon positions) and GC3 (GC content at the third codon position). A strong GC12-GC3 association is generally interpreted as evidence consistent with a predominant role of mutational bias, whereas a weak association suggests stronger constraint or selection on GC12 relative to GC3 (Sharp and Li, 1987; Ioanna *et al.*, 2024). For each species, GC12 and GC3 were computed for all genes and evaluated using regression analyses in neutrality plots.

Statistical analyses and visualization

All statistical analyses were conducted in R. Data visualization was performed using ggplot2 and pheatmap, among other packages. To assess codon usage differences between high- and low-expression gene sets, we applied t-tests and Mann-Whitney U tests, as appropriate. Differences among major lineages were evaluated using ANOVA. Multiple testing was controlled using false discovery rate (FDR) correction following the Benjamini-Hochberg procedure.

RESULTS AND DISCUSSION

Basic overview

We quantified codon-usage and protein-composition indices across nuclear CDSs from 30 angiosperm species, with emphasis on Fabaceae and comparisons with Poaceae, Brassicaceae and other outgroups. Substantial interspecific variation was detected in GC content, codon adaptation and bias indices (CAI, CBI), Fop, mean protein length (L_aa), hydropathy (GRAVY) and aromaticity (Aromo) (Fig 1).

Fabaceae generally showed genome-wide GC values of ~0.44-0.45 (*e.g.*, *Glycine max* = 0.4407; *Cajanus cajan* = 0.4545), comparable to or slightly higher than several non-legumes (*e.g.*, *Solanum lycopersicum* = 0.417) and within the range of *Arabidopsis thaliana* (0.452) (Šmarda *et al.*, 2014). At synonymous third positions, Fabaceae taxa such as *Medicago truncatula* (GC3s = 0.4522) and *Pisum sativum* (0.4419) exceeded *A. thaliana* (0.4304) and, more prominently, *S. lycopersicum* (0.3477), indicating lineage-specific divergence in third-position nucleotide composition and codon-ending preference.

CAI values in Fabaceae were concentrated at ~0.18-0.22 and were broadly comparable to those in major grasses (e.g., *Zea mays* = 0.221; *Oryza sativa* = 0.218), suggesting only moderate among-lineage differences in codon "optimization," with no uniformly dominant clade. CBI values were similar across several taxa (*e.g.*, *M. sativa* = 0.3819; *G. max* = 0.3844; *A. thaliana* = 0.3873), whereas S. *lycopersicum* was lower (0.3701), indicating that codon-bias strength remains heterogeneous even within eudicots.

Fabaceae also tended to exhibit longer mean protein lengths (L_aa ~300–370 aa; *e.g.*, *M. truncatula* = 357; *G. max* = 373), consistent with lineage-level differences in average CDS length. GRAVY values were uniformly negative, indicating an overall hydrophilic proteome tendency; for example, *M. truncatula* (-0.2003) was less negative than *O. sativa* (-0.264), consistent with a relatively more hydrophilic mean protein composition in the latter. Aromo values were moderate and relatively stable in Fabaceae (*e.g.*, *G. max* = 0.0847; *M. sativa* = 0.0839) but lower in *Z. mays* (0.0729), indicating lineage-dependent differences in amino-acid composition.

These results highlight lineage-specific divergence in base composition, particularly at third codon positions and codon preferences. This framework enhances understanding of mutation–selection dynamics and supports codon-aware gene design (Parvathy *et al.*, 2022).

ENc–GC3s analysis (ENC-plot)

In this study, we examined the relationship between the effective number of codons (ENc, also denoted Nc) and GC3s (GC content at third codon positions) using ENc-GC3s plots (Fig 2). Across most species, ENc displayed the characteristic dependence on GC3s expected under the Wright model, with genes approaching the theoretical expectation curve over a broad GC3s range. In general, genes with lower GC3s tended to exhibit higher ENc values, consistent with weaker apparent codon bias, whereas ENc decreased as GC3s increased, indicating progressively stronger skew in synonymous codon usage as GC-ending codons became more prevalent.

Within Fabaceae, species such as *Glycine max, Medicago sativa* and *Cicer arietinum* showed conspicuous clustering of genes toward lower ENc values within intermediate GC3s ranges, consistent with enhanced codon bias in GC3s-enriched genes. By contrast, several monocots (*e.g.*, *Oryza sativa* and *Zea mays*) exhibited

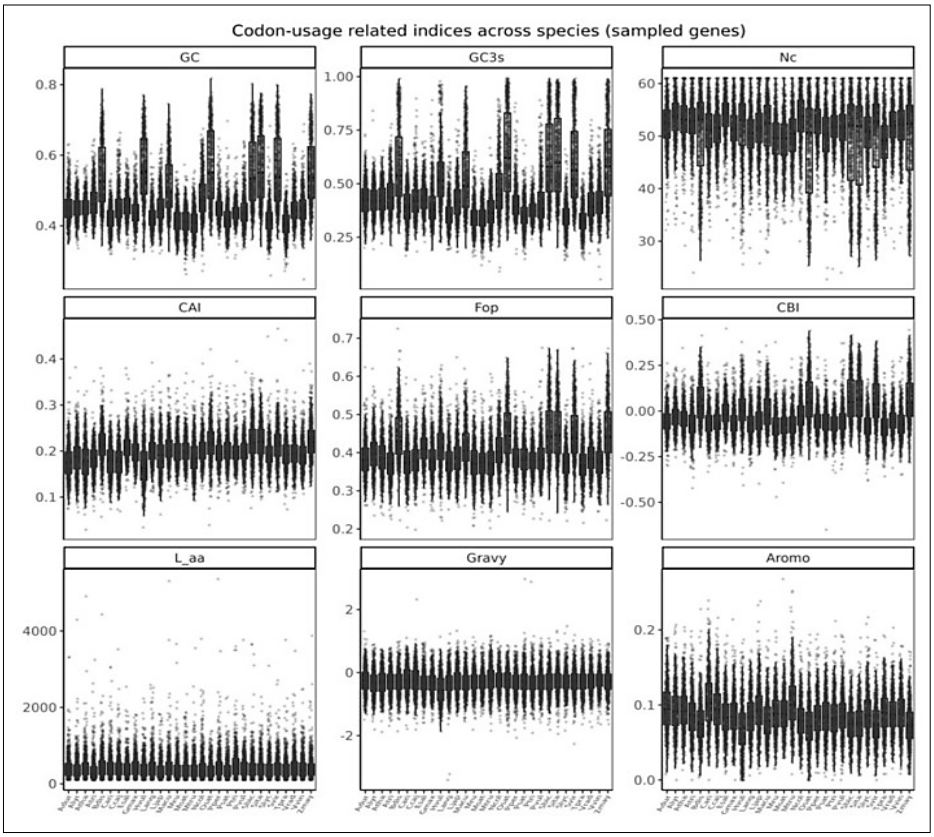


Fig 1: Basic index analysis.

comparatively stable ENc distributions across GC3s, suggesting that the GC3s dependence of ENc is less pronounced in these taxa. Together, these ENc–GC3s patterns indicate that third-position base composition constitutes a major axis underlying codon bias across angiosperms, while the magnitude and distribution of deviation from mutation-only expectations vary among lineages, implying the potential contribution of additional forces beyond base-composition constraints (Wright, 1990; Parvathy *et al*., 2022).

Importantly, the ENc-GC3s framework provides a quantitative, genome-wide diagnostic to compare codon-usage architectures among plant lineages and to motivate subsequent mechanistic inference regarding the relative roles of compositional bias and selection. The lineage-specific differences observed here further underscore that codon bias evolution is not uniform across angiosperms and highlight Fabaceae as a clade in which codon-usage structure merits deeper evaluation in conjunction with expression- and translation-related evidence.

PR2 analysis

In this study, we investigated the relationship between PR2 bias at third codon positions [A3/(A3 + A3 + T3)] and GC content across species, as depicted in the scatter plot in Fig 3. Each data point reflects the codon usage bias of sample genes within a species, with the x-axis representing G3/(G3 + C3) and the y-axis representing A3/(A3 + A3 + T3). These plots reveal biases in codon selection and the distribution patterns of specific codons within the genome. Most species' data points are concentrated in the 0.25 to 0.75 range, suggesting a consistent base distribution at third codon positions. Notably, in Fabaceae species, such as *Glycine max* (Gmax) and *Medicago truncatula* (Mtru), the scatter plots exhibit pronounced biases, indicating strong selective pressure favoring the use of adenine (A) at the third codon position. In contrast, non-legume species, such as *Arabidopsis thaliana* (Atha) and *Solanum lycopersicum* (Slyc), show more dispersed distributions, reflecting weaker codon selectivity and more random codon usage patterns.

Further analysis revealed that, Fabaceae species, such as *Glycine max* and *Cajanus cajan* (Ccaj), generally have higher A3/(A3 + A3 + T3) values, indicating stronger bias towards adenine at the third codon position. In contrast, non-legume species, including *Arabidopsis thaliana* and *Solanum lycopersicum*, exhibit lower frequencies of A3 usage, reflecting weaker codon selectivity. For monocots like *Oryza sativa* (Osat) and *Zea mays* (Zmay), their scatter plots show a significant relationship between GC content and PR2 bias. In regions of higher GC content, there is a stronger bias in the use of codons at third positions, while lower GC regions exhibit weaker bias. This suggests that higher GC content may be associated

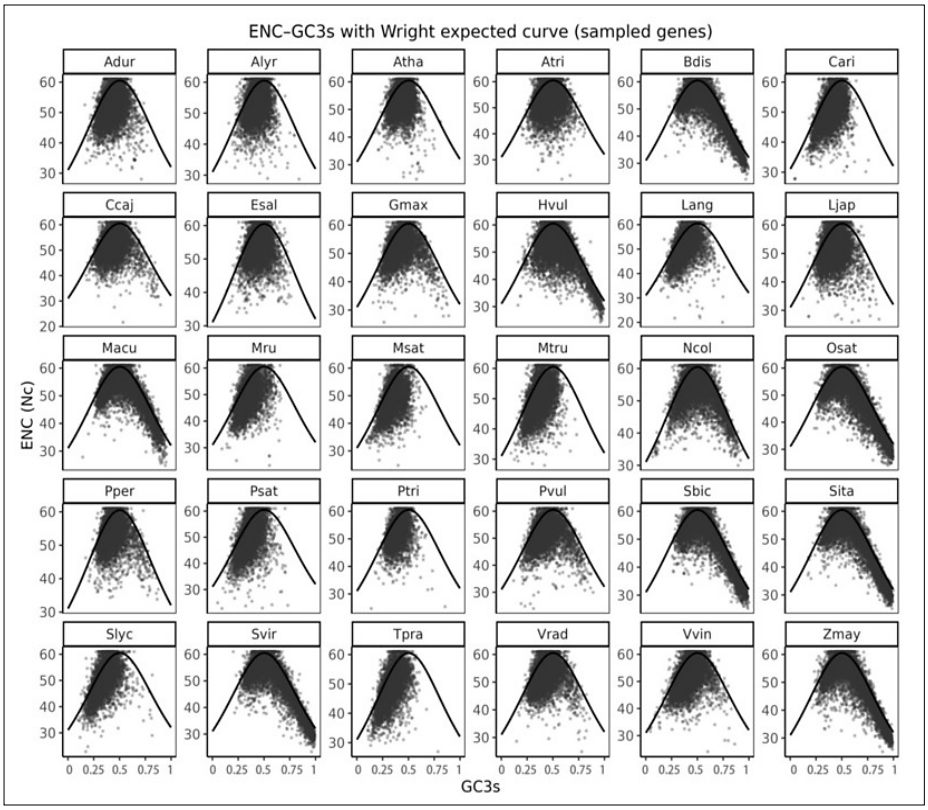


Fig 2: ENC-GC3s plot analysis of 30 species of angiosperms.

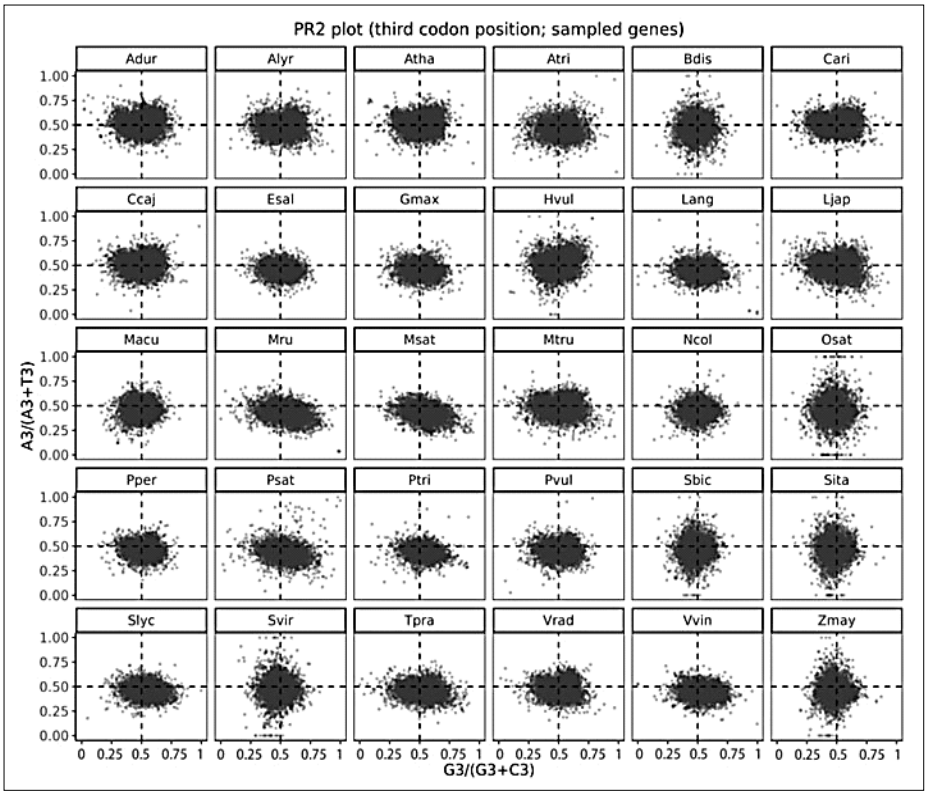


Fig 3: Scatter plot indicating the inter-relationship between PR2 bias at third codon positions and GC content across 30 species of angiosperms.

with stronger selective pressures, leading to more pronounced codon usage bias.

Overall, the relationship between PR2 bias and GC content highlights significant differences in codon selection and gene expression regulation across species. Fabaceae typically exhibits strong codon bias, while other plant lineages display broader, more diverse codon usage patterns. These differences are likely tied to the evolutionary adaptive pressures and gene expression regulatory mechanisms faced by each lineage, providing important theoretical insights for future plant genome research and crop improvement (Yang *et al.*, 2023; Parvathy *et al.*, 2022).

Correlation analysis

Fig 4 presents a Spearman correlation heatmap of six codon usage indices: GC, GC3s, Nc (effective number of codons), CAI (Codon Adaptation Index), CBI (Codon Bias Index) and Fop (Frequency of Optimal Codons). The correlation coefficients (rho values) range from -1 to 1, representing the strength and direction of the linear relationships between these indices. The depth of color reflects the strength of the correlation, with darker shades indicating stronger positive correlations and lighter shades indicating weaker or no correlation.

From the heatmap, it is evident that GC and GC3s exhibit a very strong positive correlation (rho = 0.82), suggesting a close relationship between the GC content at third codon positions and the overall GC content of the genome. This finding aligns with the synchronous variation of GC content during genome evolution (Bowers *et al.*, 2022). Additionally, strong positive correlations were observed between GC and both Fop (rho = 0.61) and CBI (rho = 0.65), indicating that species with higher GC content tend to exhibit stronger codon selectivity, which contributes to improved gene expression efficiency (Hao *et al.*, 2025).

The correlation between CAI and other indices showed a notable positive relationship with Fop (rho = 0.66) and CBI (rho = 0.55), while no significant correlation was found with Nc (rho = -0.11), suggesting that CAI primarily reflects the degree of codon optimization, whereas Nc is more closely related to broader genomic characteristics (Kwon *et al.*, 2016). Furthermore, GC3s also exhibited moderate positive correlations with Fop (rho = 0.55) and CBI (rho = 0.64), further supporting the idea that changes in GC3s are tightly linked to codon selection, particularly at third codon positions, where an increase in GC content typically accompanies codon usage optimization (Ruden, 2025).

Overall, the correlation results from the heatmap provide a deep understanding of the interrelationships between different codon usage indices, revealing the complex interactions between GC content, codon optimization and gene expression efficiency. These findings not only validate the relationship between codon selection and genomic features but also offer a theoretical foundation for further studies on the mechanisms of gene expression regulation.

RSCU analysis

Fig 5 presents an RSCU (Relative Synonymous Codon Usage) heatmap of the species × codon matrix, revealing codon usage preferences across different species. The RSCU values were normalized so that each codons value is centered around 1. The color intensity in the heatmap

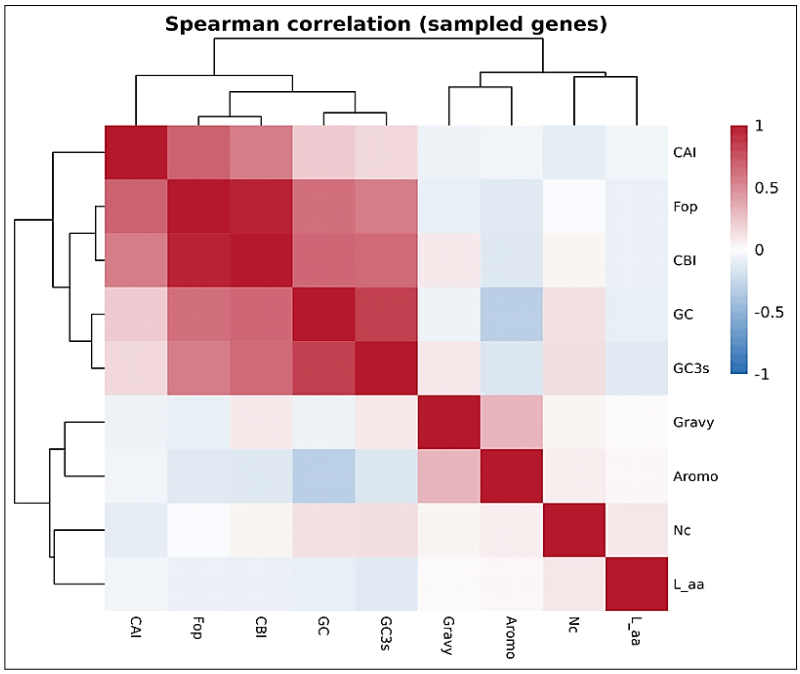


Fig 4: Spearman correlation heatmap of six codon usage indices.

represents the frequency of codon usage in each species, with deep red indicating a strong preference for a specific codon and deep blue indicating a bias toward the use of another codon. These data were processed using hierarchical clustering, with the x-axis representing codon types and the y-axis representing species, thereby illustrating the codon usage bias patterns across species. From the heatmap, significant differences in codon selection were observed among different plant lineages (*e.g.*, Fabaceae, Poaceae, monocots). Fabaceae species, such as *Glycine max* and *Medicago truncatula*, along with Poaceae species like *Zea mays* and *Oryza sativa*, exhibited consistent RSCU patterns for specific codons, indicating strong selection for the usage of certain codons within their genomes. For example, *Glycine max* (Gmax) and *Medicago truncatula* (Mtru) demonstrated higher RSCU values for codons associated with GC preference, suggesting these species have optimized the use of high-GC codons for gene expression. These codons may be associated with plant adaptability and environmental stress responses, particularly in enhancing gene expression efficiency and stability (Parvathy *et al.*, 2022).

In contrast, monocots such as *Oryza sativa* (Osat) and *Zea mays* (Zmay) also showed some optimization in codon usage, especially for codons in regions of their genomes with high GC content. These species exhibited different codon usage patterns compared to Fabaceae, likely reflecting their distinct evolutionary histories and genomic adaptations. The variation in codon usage patterns across different lineages further supports the notion that codon usage bias reflects differences in gene expression regulation, environmental adaptation and genomic optimization strategies (Liu *et al.*, 2004). Overall, the RSCU heatmap provides a visual representation of codon usage preferences across species, shedding light on the selection and optimization of codons in plant genomes. These findings offer valuable insights for understanding the mechanisms of codon selection in plants and provide a useful reference for plant genome improvement and crop optimization.

Fig 6 presents a principal component analysis (PCA) of species based on RSCU (Relative Synonymous Codon Usage) features. Using PCA, we projected the codon usage preferences of species into a two-dimensional space, where PC1 and PC2 explained 77.4% and 10.3% of the total data variance, respectively. Each point in the plot represents a species, with species grouped according to their codon usage characteristics and the shapes and colors of the points indicate the species' lineage. From the plot, it is apparent that species are clustered into distinct groups based on codon usage preferences. Notably, along PC1, there is significant separation between Fabaceae species (*e.g.*, *Glycine max* and *Medicago truncatula*) and Poaceae species (*e.g.*, *Zea mays* and *Oryza sativa*), suggesting substantial differences in codon usage between these two groups. Fabaceae species predominantly cluster in the positive range of PC1, while Poaceae species are concentrated in the negative range,

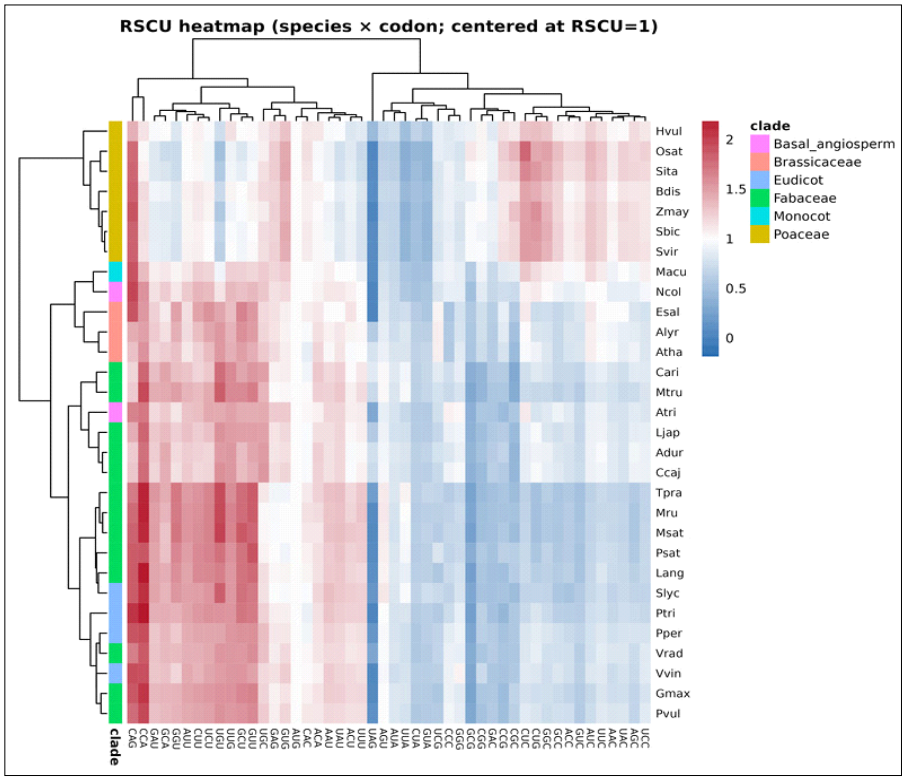


Fig 5: Relative Synonymous Codon Usage heatmap of 30 angiosperms species × codon matrix.

reflecting distinct preferences for codon selection at the genomic level. Along PC2, differences in codon usage between monocots and dicots are also evident, with species from the Brassicaceae family (*e.g.*, *Arabidopsis thaliana*, *Brassica oleracea*) showing a different distribution pattern compared to monocots like *Oryza sativa* and *Zea mays*. This variation may be related to differences in gene expression optimization strategies between these plant groups (Yang *et al.*, 2023). Furthermore, the PCA clustering analysis revealed distinct differences between Basal angiosperms (*e.g.*, *Amborella trichopoda*) and Eudicots (*e.g.*, *Populus trichocarpa*), highlighting differences in their genomic codon usage patterns, closely tied to their respective evolutionary histories and adaptive traits. Overall, the RSCU PCA plot demonstrates significant differences in codon preferences across species, providing visual evidence to support the understanding of codon usage and gene expression optimization mechanisms between different plant lineages. These results offer critical insights into the gene expression regulation and adaptive evolution mechanisms across species (Majeed *et al.*, 2026).

Neutrality plot analysis

Fig 7 shows the relationship between GC12 and GC3 and was used to infer the relative contributions of mutational bias and selective constraints to genome-wide base composition across species. For each species, we fitted a linear regression between GC12 and GC3 and summarized the association using the slope, R^2 and p value. Here, the slope quantifies the extent to which variation at third codon positions is mirrored at the first and second positions, whereas R^2 reflects the goodness-of-fit and the p value evaluates statistical significance.

Across species, many regression fits exhibited significant positive associations, although the strength of the relationship varied substantially among lineages. In several Fabaceae species, GC12 and GC3 were positively associated; for example, *Glycine max* showed a measurable relationship (R^2 = 0.285; $p < 1\times10^{-300}$), indicating coordinated variation in GC content across codon positions. Such coupling is consistent with non-random constraints on compositional architecture and has been interpreted as reflecting the combined action of mutation and selection shaping codon-position–specific base composition (Wang *et al.*, 2025). In contrast, Medicago truncatula displayed a much weaker relationship (R^2 = 0.002), suggesting that the linkage between GC3 and GC12 is minimal in this species. Similarly, taxa such as *Arabidopsis thaliana* showed weak coupling (*e.g.*, R^2 = 0.005), implying that compositional variation at third positions is only weakly reflected at the more constrained first and second positions in these genomes.

In addition, several species exhibited comparatively strong GC12-GC3 coupling; for instance, *Zea mays* showed a robust positive association (R^2 = 0.409), consistent with more pronounced genome-wide coordination of base composition across codon positions. Collectively, the cross-species variation in slope and R^2 values highlights substantial heterogeneity in codon-position compositional coupling among angiosperms. This neutrality-plot framework provides an interpretable quantitative summary of how base composition at third positions relates to that at the more functionally constrained first and second positions and thus offers a useful basis for comparing mutational and selective influences on

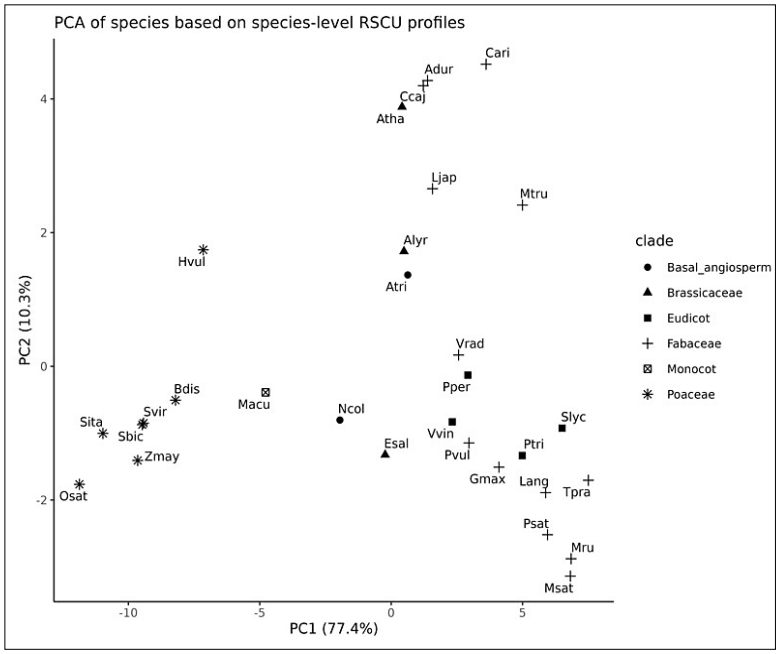


Fig 6: Principal component analysis of species based on Relative Synonymous Codon Usage features across 30 angiosperm species.

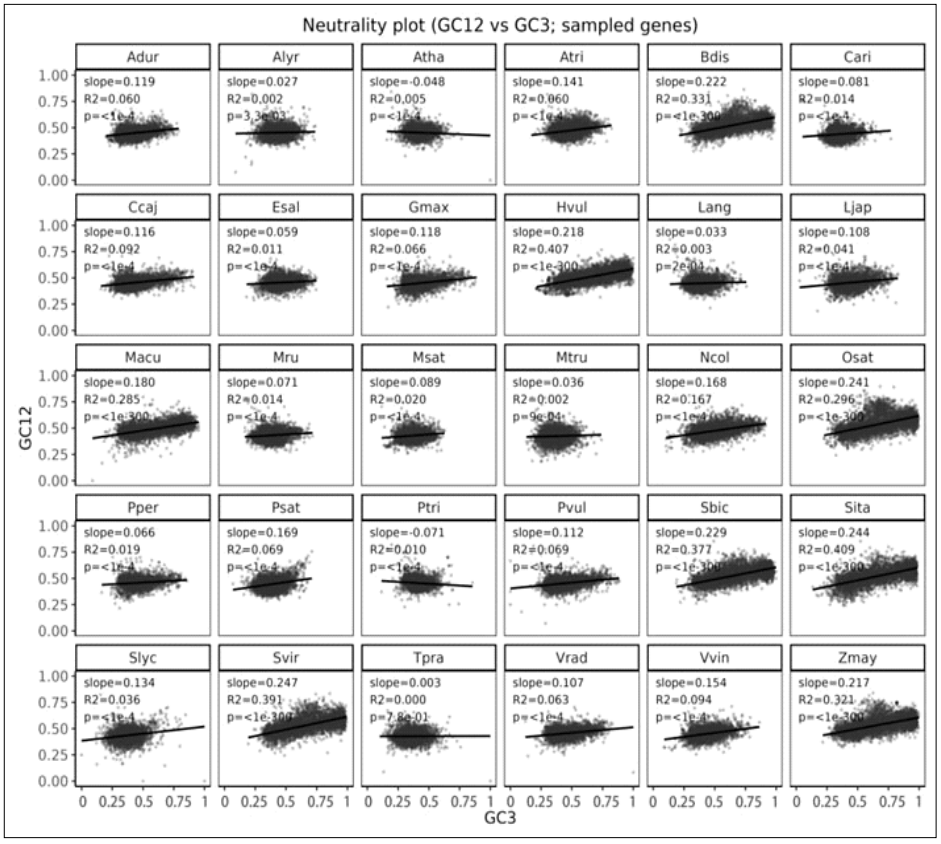


Fig 7: Neutrality plot analysis indicating the relationship between GC12 and GC3.

compositional evolution among lineages (Alemu *et al.*, 2024; Glémin *et al.*, 2014).

RSCU-based clustering analysis

Fig 8 shows a species clustering tree constructed from RSCU (relative synonymous codon usage) profiles using the neighbor-joining (NJ) method. This tree groups species according to the similarity of their codon usage patterns, thereby providing an integrative view of how compositional constraints and putative selective effects shape codon preference architectures across taxa. *Amborella trichopoda* (Atri) was used as an outgroup to root the tree, facilitating interpretation of lineage-level relationships.

The topology revealed clear lineage-associated structure, with major families (*e.g.*, Fabaceae, Poaceae and Brassicaceae) forming recognizable clusters, consistent with the presence of phylogenetic signals in codon usage patterns. For example, Fabaceae species such as *Glycine max* and *Medicago truncatula* clustered closely, indicating broadly similar codon usage profiles within the family. In contrast, Poaceae species (*e.g.*, *Zea mays* and *Sorghum bicolor*) occupied distinct branches, suggesting a codon usage architecture that differs systematically from that of legumes and other eudicot lineages, potentially reflecting lineage-specific compositional backgrounds and evolutionary histories (Wang *et al.*, 2025).

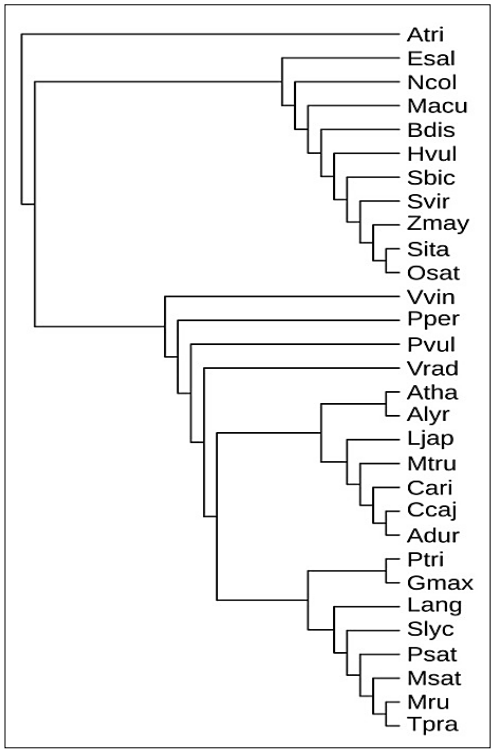


Fig 8: RSCU-based clustering analysis of 30 angiosperm species.

More broadly, the distribution of taxa across the NJ tree highlights substantial diversity in codon usage patterns among plant families. The separation between Poaceae and Fabaceae, for instance, is consistent with their divergent base-composition regimes and codon-ending preferences, which may in turn be associated with differences in genome organization and gene expression-related constraints. Collectively, the RSCU-based clustering provides a compact, species-level summary of codon preference similarity, offering an additional line of evidence that codon usage evolution varies across angiosperm lineages and is shaped by a combination of compositional bias and lineage-dependent selective constraints (Glémin *et al*., 2014; Alemu *et al*., 2024).

CONCLUSION

Centered on Fabaceae and incorporating Poaceae, Brassicaceae and other outgroup taxa, this study performed a systematic comparison of codon usage patterns in nuclear genome CDSs from 30 angiosperm species under a unified data-cleaning and computational pipeline. We quantified GC/GC3s, ENc (Nc) and expression-related indices (CAI/CBI/Fop) and further constructed a species-level RSCU matrix to delineate cross-lineage codon preference landscapes. Correlation analyses indicated that third-position base composition constitutes a principal axis shaping the global preference structure: GC was highly concordant with GC3s (ρ = 0.82) and showed strong positive associations with CBI and Fop (ρ = 0.65 and 0.61), supporting coordinated shifts between compositional bias and bias strength across species. RSCU-based PCA corroborated this structured differentiation, with PC1 alone explaining 77.4% of the variance and clearly separating Fabaceae from Poaceae in codon preference space. Meanwhile, neutrality-plot analyses revealed that the relative contributions of mutational bias and selection differ across lineages (*e.g.*, stronger GC12-GC3 coupling in *Z. mays* and *G. max* but weaker coupling in *M. truncatula* and *A. thaliana*), implying lineage- and genomic-background–dependent selective signatures superimposed on compositional constraints.

Collectively, this work establishes a cross-species baseline of codon preference for legume crops and forage species, providing a reusable reference for codon recoding, heterologous expression optimization and sequence design in molecular breeding. Future studies should integrate transcriptomic, ribosome profiling and tRNA supply information and apply phylogenetically controlled models to disentangle lineage signals from compositional effects, thereby more precisely quantifying the strength and targets of selection acting on codon usage.

ACKNOWLEDGMENT

This work was supported by the National Grassland Technology Innovation Center (Preparatory) Project (Grant No. CCPTZX2024GJ09) and the Graduate Scientific Research Innovation Program of the Inner Mongolia Autonomous Region (Grant No. KC2025029S).

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.

Conflict of interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

REFERENCES

Alemu, A., Åstrand, J., Montesinos-López, O.A., Sánchez, J.I., González, J.F., Tadesse, W., Vetukuri, R.R., Carlsson, A.S., Ceplitis, A., Crossa, J., Ortiz, R. and Chawade, A. (2024). Genomic selection in plant breeding: Key factors shaping two decades of progress. *Molecular Plant.* **17(4):** 552-578. doi: 10.1016/j.molp.2024.03.007.

Basudha, Ch., Chanu, K.V., Sobita, N. and Ningombam, A. (2025). The complete mitochondrial genome of *Bangana dero* (Cyprinidae: Labeoninae) and its relationship with other labeonin fishes. *Indian Journal of Animal Research.* **59(12):** 2018-2027. doi: 10.18805/IJAR.B-5476.

Bertioli, D.J., Cannon, S.B., Froenicke, L., Huang, G., Farmer, A.D., Cannon, E.K., Liu, X. *et al.* (2016). The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nature Genetics*. **48:** 438-446.

Bowers, J.E., Tang, H., Burke, J.M. and Paterson, A.H. (2022). GC content of plant genes is linked to past gene duplications. *Plos One*. doi: 10.1371/journal.pone.0261748.

Cheeran, K., Suresh, K.P., Jacob, S.S., Sathish Gowda, C.S., Gejendiran, N., Sridevi, R. and Patil, S.S. (2024). Analysis of codon usage bias of six genes of replicase/coat protein of Tobacco mosaic virus. *Indian Journal of Agricultural Research.* **58(4):** 720-726. doi: 10.18805/IJARe.A-6107.

Glémin, S., Clément, Y., David, J. and Ressayre, A. (2014). GC content evolution in coding regions of angiosperm genomes: A unifying hypothesis. *Trends in Genetics.* **30(7):** 263-270. doi: 10.1016/j.tig.2014.05.002.

Grantham, R., Gautier, C., Gouy, M., Mercier, R. and Pavé, A. (1980). Codon catalog usage and the genome hypothesis. *Nucleic Acids Research*. **8:** 49-62.

Hanson, G. and Coller, J. (2018). Codon optimality, bias and usage in translation and mRNA decay. *Nature Reviews Molecular Cell Biology*. **19(1):** 20-30.

Hao, J., Liang, Y., Wang, T. and Su, Y. (2025). Correlations of gene expression, codon usage bias and evolutionary rates of the mitochondrial genome show tissue differentiation in *Ophioglossum vulgatum*. *BMC Plant Biology.* **25(1):** 134. doi: 10.1186/s12870-025-06157-x.

Ioanna, K., Carolin, K. and Rui, B. (2024). The patterns of codon usage between chordates and arthropods are different but co-evolving with mutational biases. *Molecular Biology and Evolution.* **41(5):** msae080.

Kawabe, A. and Miyashita, N.T. (2003). Patterns of codon usage bias in three dicot and four monocot plant species. *Genes and Genetic Systems.* **78(4):** 343-352.

Kwon, K.-C., Chan, H.-T., León, I.R., Williams-Carrier, R., Barkan, A. and Daniell, H. (2016). Codon optimization to enhance expression yields insights into chloroplast translation. *Plant Physiology.* **172(1):** 62-77. doi: 10.1104/pp.16.00981.

Liu, Q., Feng, Y., Zhao, X., Dong, H. and Xue, Q. (2004). Synonymous codon usage bias in *Oryza sativa*. *Plant Science.* **167(1):** 101-105. doi: 10.1016/j.plantsci.2004.03.003.

Majeed, A., Sharma, V., Rehman, W.U., Kaur, A., Das, S., Joseph, J., Singh, A. and Bhardwaj, P. (2026). Comprehensive codon usage analysis across diverse plant lineages. *Biochemical Genetics*. **64:** 727-750. doi: 10.1007/s10528-025-11053-y.

Parvathy, S.T., Udayasuriyan, V. and Bhadana, V. (2022). Codon usage bias. *Molecular Biology Reports*. **49(1):** 539-565. doi: 10.1007/s11033-021-06749-4.

Peden, J.F. (1999). Analysis of codon usage. PhD Thesis, University of Nottingham.

Plotkin, J.B. and Kudla, G. (2011). Synonymous but not the same: The causes and consequences of codon bias. *Nature Reviews Genetics*. **12(1):** 32-42.

Presnyak, V., Alhusaini, N., Chen, Y.-H., Martin, S., Morris, N., Kline, N., Olson, S., Weinberg, D., Baker, K.E., Graveley, B.R. and Coller, J. (2015). Codon optimality is a major determinant of mRNA stability. *Cell.* **160(6):** 1111-1124.

Ruden, D.M. (2025). GC content in nuclear-encoded genes and effective number of codons (ENC) are positively correlated in AT-rich species and negatively correlated in GC-rich species. *Genes (Basel)*. **16(4):** 432. doi: 10.3390/genes16040432.

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., Kato, T., Nakao, M., Sasamoto, S. *et al.* (2008). Genome structure of the legume, *Lotus japonicus*. *DNA Research*. **15(4):** 227-239.

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W. *et al.* (2010). Genome sequence of the palaeopolyploid soybean. *Nature.* **463:** 178-183.

Sharp, P.M. and Li, W.-H. (1987). The codon adaptation index-A measure of directional synonymous codon usage bias and its potential applications. *Nucleic Acids Research.* **15(3):** 1281-1295.

Šmarda, P., Bureš, P., Horová, L., Leitch, I. J., Mucina, L., Pacini, E., Tichı, L., Grulich, V. and Rotreklová, O. (2014). Ecological and evolutionary significance of genomic GC content diversity in monocots. *Proceedings of the National Academy of Sciences of the United States of America*. **111(39):** E4096-E4102. doi: 10.1073/pnas.1321152111. Epub 2014 Sep 15. PMID: 25225383; PMCID: PMC4191780.

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. (1992). Directional mutation pressure, selective constraints and genetic equilibria. *Journal of Molecular Evolution.* **34:** 95-114.

Sueoka, N. (1995). Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *Journal of Molecular Evolution*. **40(3):** 318-325.

Tatarinova, T.V., Alexandrov, N.N., Bouck, J.B. and Feldmann, K.A. (2010). GC3 biology in corn, rice, sorghum and other grasses. *BMC Genomics*. **11:** 308.

Varshney, R.K., Song, C., Saxena, R.K., Azam, S., Yu, S., Sharpe, A.G., Cannon, S. *et al.* (2013). Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nature Biotechnology*. **31:** 240-246.

Wang, L., Jiang, X., Jiao, W., Mao, J., Ye, W., Cao, Y., Chen, Q. and Song, Q. (2025). Pangenome analysis provides insights into legume evolution and breeding. *Nature Genetics.* **57:** 2052-2061. doi: 10.1038/s41588-025-02280-5.

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

Yang, Q., Xin, C., Xiao, Q.-S., Lin, Y.-T., Li, L. and Zhao, J.-L. (2023). Codon usage bias in chloroplast genes implicates adaptive evolution of four ginger species. *Frontiers in Plant Science.* **14:** 1304264. doi: 10.3389/fpls.2023.1304264.

Yang, Q., Xin, C., Xiao, Q.-S., Lin, Y.-T., Li, L. and Zhao, J.-L. (2023). Codon usage bias in chloroplast genes implicates adaptive evolution of four ginger species. *Frontiers in Plant Science.* **14:** 1304264. doi: 10.3389/fpls.2023.1304264.

Young, N.D., Debellé, F., Oldroyd, G.E., Geurts, R., Cannon, S.B., Udvardi, M.K., Benedito, V.A. *et al.* (2011). The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature.* **480:** 520-524.

Zhang, J., Mu, Y., Zhu, W. and Yang, X. (2025). A rare mitochondrial genome of Albino *Eothenomys eleusis* Thomas 1911 (Cricetidae: Arvicolinae) from southeastern Yunnan, China and its phylogenetic analysis. *Indian Journal of Animal Research*. **59(4):** 560-567. doi: 10.18805/IJAR.BF-1882.