Theme: Legume-Based Solutions for Nutritional Security and Sustainable Agro-Ecosystems

RESEARCH ARTICLE Legume Research- An International Journal, Volume 49 Special Issue: 10-20 (2026)

Comparative Analysis of Codon Usage Bias in bHLH Genes Across 30 Angiosperm Species Centered on Fabaceae

Yuehua Zhang[1], Jingshi Lu[1], Fang Tang[1], Fengling Shi[1]

10.18805/LRF-934

ABSTRACT

Background: Synonymous codon usage bias (CUB) is a pervasive feature of plant nuclear genomes, shaped by lineage-specific base composition and additional evolutionary forces. Despite its well-documented role in genome evolution, the codon-preference landscape of Fabaceae relative to non-Fabaceae lineages remains unclear. This study establishes a comparative framework across 30 angiosperm species, centered on Fabaceae, to delineate lineage-scale differentiation patterns and identify key determinants of synonymous codon usage.

Methods: A total of 5,434 bHLH genes were identified from 30 species and codon-bias profiling was performed under standardized quality-control protocols. Key indices of codon bias, including GC3s, the effective number of codons (ENc/Nc), codon adaptation index (CAI), codon bias index (CBI) and frequency of optimal codons (Fop), were calculated. ENc-GC3s analyses were integrated with the wright expected curve and strand-asymmetric base usage at the third codon position was examined using PR2-bias metrics. Multivariate techniques, such as hierarchical clustering, principal component analysis (PCA) and correspondence analysis (CA), were employed to validate global codon-preference structures and assess lineage-specific variability.

Result: Species-level analyses revealed consistent phylogenetic stratification in codon-preference strength and third-position base composition. Poaceae exhibited higher GC3s (up to 0.764) and lower Nc (minimum 42.32), whereas Fabaceae species clustered within a lower GC3s range (~0.32-0.44) with differentiated Nc values (47-52), establishing GC3s as a dominant factor influencing codon-preference divergence. Despite adherence to the Wright expectation, deltaENC medians were predominantly negative, indicating a “stronger-than-expected” codon bias. PR2 analysis identified substantial interspecific differences in third-position base asymmetry, with distinct directional biases in A/T and G/C usage in Fabaceae and Poaceae. Multivariate analysis confirmed these findings, with PCA explaining 86.5% of the variance along the primary axis of codon-preference, and CA independently validating species clustering.

Key words: bHLH, Codon usage bias, ENc, Fabaceae, GC3s.

INTRODUCTION

Synonymous codon usage is a non-random process that exhibits quantifiable variation among species, genomes and gene families. Codon usage bias is influenced by mutational pressure, lineage-specific base composition and additional evolutionary and functional factors, including translational efficiency, co-translational protein folding and the regulation of mRNA stability, all of which may affect molecular evolution and gene expression (Bulmer, 1991; Drummond and Wilke, 2008; Plotkin and Kudla, 2011). Notably, codon optimality can modulate mRNA stability and decay, indicating that synonymous substitutions may alter expression and regulatory efficiency without changing the encoded amino acid sequence (Presnyak *et al*., 2015; Hanson and Coller, 2018). Understanding codon usage variation across species is therefore important for investigating coding-sequence evolution and may also inform the rational design of genes for heterologous expression and crop biotechnology (Sharp and Li, 1987; Plotkin and Kudla, 2011).

In plants, transcription-factor families such as the basic helix-loop-helix (bHLH) family combine conserved molecular functions with lineage-associated variation in gene number and functional diversification. bHLH transcription factors participate in numerous developmental, metabolic and stress-response pathways and have diversified extensively during angiosperm evolution (Bailey *et al*., 2003). Fabaceae is a major agricultural and ecological lineage that includes important forage, grain and oil crops, such as alfalfa, soybean and pea and several representative legumes possess distinctive genomic histories involving duplication and polyploidization (Schmutz *et al*., 2010).

Recent studies have characterized codon-usage variation in Fabaceae at different analytical scales. A genome-

[1]College of Grassland Science, Inner Mongolia Agricultural University, Hohhot, 010010, China.

Corresponding Author: Fengling Shi, College of Grassland Science, Inner Mongolia Agricultural University, Hohhot, 010010, China. Email: sfl0000@126.com

How to cite this article: Zhang, Y., Lu, J., Tang, F. and Shi, F. (2026). Comparative Analysis of Codon Usage Bias in bHLH Genes Across 30 Angiosperm Species Centered on Fabaceae. *Legume Research*. **49 (Special Issue):** 10-20. doi: 10.18805/LRF-934.

Submitted: 30-01-2026 **Accepted:** 16-05-2026 **Online:** 30-06-2026

wide comparison of nuclear coding sequences across 30 angiosperm species established a broad Fabaceae-centered reference for variation in GC3s, ENc, RSCU and related codon-usage indices (Lu *et al*., 2026). At the single-genus level, transcriptome-based analyses of Medicago ruthenica and Medicago varia reported preferences for codons ending in A or U, with a mean GC3s of 0.33 reported for the M. ruthenica transcriptome (Peng *et al*., 2025; Bai *et al*., 2025). However, genome-wide and transcriptome-wide patterns cannot necessarily be extrapolated to a functionally defined transcription-factor family. It therefore remains unclear whether bHLH genes recapitulate the broad codon-usage structure observed across whole coding-sequence datasets or exhibit gene-family-specific variation.

Cross-species codon-usage analysis also presents several methodological challenges. First, variation in GC content at synonymous third codon positions can account for a large proportion of the observed codon-usage variation, complicating the separation of compositional effects from other evolutionary influences. Second, alternative representations of codon usage, including RSCU profiles, codon-count matrices and ENc-GC3s relationships, differ in scale, weighting and interpretation. Combinations of RSCU analysis, neutrality plots and parity rule 2 plots have also been applied in recent codon-usage studies, including analyses of plant-associated viral genes (Cheeran *et al*., 2024). Integrating complementary indices and multivariate approaches can therefore provide a more comprehensive description of codon-usage structure and permit assessment of cross-method consistency. Nevertheless, agreement among these analyses does not, by itself, establish the relative contributions of mutation, natural selection, translational constraints, or other evolutionary processes (Bulmer, 1991; Sharp and Li, 1987; Plotkin and Kudla, 2011).

Building on the available genome-wide and transcriptome-wide evidence, the present study applies a standardized gene-family-specific workflow to bHLH coding sequences from 30 angiosperm species, with Fabaceae as the focal lineage. We aimed to characterize variation in codon-bias strength and synonymous third-position composition, determine whether bHLH codon-usage profiles exhibit taxonomically structured differentiation and assess whether patterns obtained from complementary codon-usage representations are concordant. Hierarchical clustering, principal component analysis and correspondence analysis were used to describe the dominant axes of species-level variation, while ENc-GC3s, PR2 and neutrality analyses were used to characterize deviations from composition-based expectations. This study provides a gene-family-specific reference for evaluating how bHLH codon usage relates to broader angiosperm patterns and establishes a basis for subsequent functional and evolutionary investigations.

MATERIALS AND METHODS

Genome data sources and sequence preprocessing

This study analyzed nuclear genome CDSs and protein sequences from 30 angiosperm species, including monocots, eudicots and Fabaceae. CDS datasets underwent quality control, including: (i) sequence length divisible by three; (ii) removal of ambiguous nucleotides or non-canonical characters; and (iii) exclusion of CDSs with internal stop codons. For genes with multiple isoforms, a single representative transcript was retained. The experiment was conducted in 2025 at Inner Mongolia Agricultural University.

Identification of gene family members

bHLH transcription factor members were identified using conserved-domain searches with HMMs. Species-specific HMM profiles were applied to protein sequences to retrieve candidates, which were then filtered for domain consistency and secondary confirmation to reduce false positives. This approach, implemented within the HMMER framework, improved sensitivity and cross-species consistency (Eddy, 2011).

Codon usage bias indices and nucleotide composition parameters

Codon usage bias indices, including GC1, GC2, GC3 and GC3s, were calculated for each CDS. ENC/Nc quantified codon bias, with lower ENC values indicating stronger preference. The ENC–GC3s curve (Wright, 1990) assessed the relationship between ENC and GC3s. Fop and CAI were calculated to evaluate codon preference, with CAI reflecting concordance with a reference preferred-codon set (Sharp and Li, 1987). These indices were computed using CodonW and/or reproducible scripts (Peden, 1999).

Construction of species-level RSCU profiles and matrices

RSCU was calculated for 59 synonymous codons (excluding Met, Trp and stop codons) across species. Species-level RSCU profiles were assembled into a matrix (59 codons × 30 species) for clustering and dimensionality reduction. A 64-codon count matrix (64 codons × 30 species) was also constructed for validation and Correspondence Analysis (CA).

PR2 bias analysis

PR2 analysis evaluated nucleotide asymmetry at the third codon position, using A3/(A3+T3) and G3/(G3+C3) to measure A/T and G/C balance. The distance to (0.5, 0.5) quantified bias intensity, enabling cross-species comparisons (Sueoka, 1995).

Neutrality plot analysis

Neutrality plots assessed the relationship between GC12 (GC1 and GC2 mean) and GC3 (or GC3s), with the regression slope indicating compositional coupling. Analyses were conducted across species for comparability.

Multivariate analyses: Clustering, PCA and correspondence analysis (CA)

Hierarchical clustering and PCA analyzed species-level RSCU data. CA, using a 64-codon count matrix, complemented RSCU-PCA for robust codon preference structure analysis (Jolliffe, 2002; Greenacre, 2007).

Data analysis workflow

All analyses were performed using a reproducible scripting workflow, ensuring consistency in parameters and computational environments. Figures and statistical summaries were generated within the same pipeline for traceability.

RESULTS AND DISCUSSION

bHLH gene family identification results

A total of 5,434 bHLH genes were identified across 30 angiosperm species, with family sizes ranging from 62 to 425 (mean = 181.13). The bHLH family is broadly conserved, though lineage-dependent expansion and contraction are evident. Most species (76.7%) had 100-249 bHLH genes, with notable deviations in *Glycine max* (425), *Zea mays* (352) and *Musa acuminata* (323) and smaller families in *Amborella trichopoda* (62) and *Nymphaea colorata* (70). Fabaceae species showed internal variation, with Gmax and Psat exhibiting large expansions, while *Medicago sativa* (105) and *Vigna radiata* (112) had smaller families. The mean family size in legumes (180.08) was similar to that in non-legumes (181.94), with the high count in *Glycine max* contributing substantially to the within-Fabaceae variation (Fig 1).

From an evolutionary perspective, the bHLH family size variation aligns with gene duplication and polyploidy events. The large family in Gmax is consistent with its paleopolyploidy background (Schmutz *et al*., 2010), while the complexity of *Zea mays* also supports family expansion (Schnable *et al*., 2009). Smaller families in atri and Ncol suggest stronger conservation or fewer amplification events. This variation provides a framework for codon-usage analysis, highlighting the need to consider species-specific effects and avoid misattribution of family-size differences to codon-preference divergence.

Cross-species divergence in core indices and distributional signatures

Descriptive statistics of CDSs from 30 angiosperms revealed significant stratification in synonymous codon usage. The effective number of codons (ENC/Nc) was low in Poaceae (*e.g*., Svir 42.323) and higher in basal lineages (*e.g*., Ncol 53.648). This divergence reflects structured codon-bias patterns, driven by base composition and mutation (Sueoka, 1988; Sueoka, 1992; Plotkin and Kudla, 2011).

GC3s exhibited a stronger contrast, with Fabaceae species (*e.g*., Mru 0.3200) clustering in a low-GC3s range and Poaceae species (*e.g*., Svir 0.7636) in a high-GC3s range (Fig 2). This contrast was further confirmed by third-position base fractions, where Poaceae showed higher C3s and G3s and Fabaceae higher A3s and T3s (Fig 2).

Fop and CAI also demonstrated consistent stratification, with Fabaceae species showing lower Fop (*e.g*., Mru 0.3669) and CAI (*e.g*., Msat 0.2043) and Poaceae higher values (*e.g*., Svir 0.5034, Sita 0.5019). These trends suggest that codon usage is influenced by compositional biases and translation-related constraints (Bulmer, 1991; Akashi, 1994; Plotkin and Kudla, 2011).

Poaceae exhibited higher gene-level heterogeneity, with larger standard deviations in GC3s and Nc, while Fabaceae showed more stable distributions (Sueoka, 1992; Plotkin and Kudla, 2011).

The contrast between Fabaceae and Poaceae, in terms of GC3s, Nc, Fop and CAI, provides a reproducible baseline for codon-bias divergence, supporting downstream multivariate analyses and enhancing the Fabaceae-centered framework (Fig 2).

ENC-GC3 analysis

Gene-level ENC-GC3s distributions from 30 species showed shifts below Wright's expected curve (Wright, 1990).

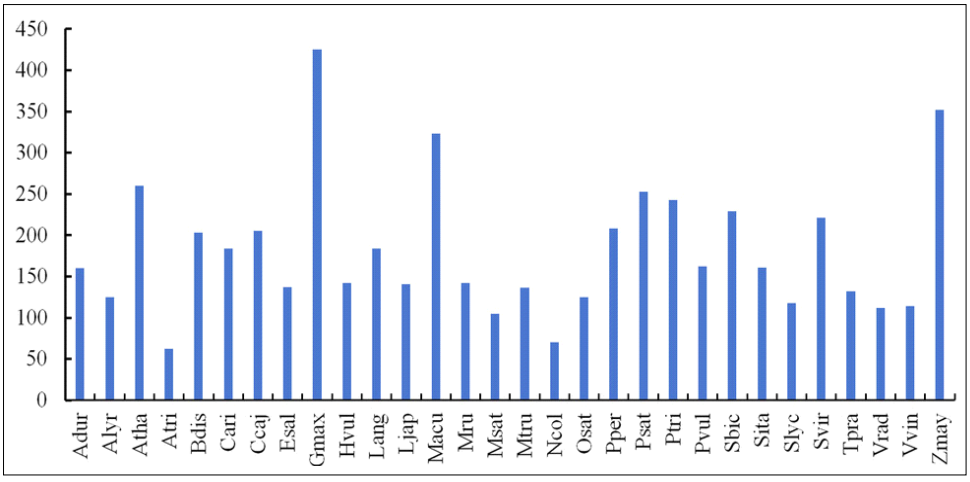


Fig 1: Identification results of the bHLH gene family across 30 angiosperm species.

ΔENC was mainly negative, indicating codon-bias strength exceeding mutational pressure (Sueoka, 1988; Hershberg and Petrov, 2008).

Poaceae species clustered in the high-GC3s region with low ENC (*e.g.*, Svir 42.32, Sita 42.53), while Fabaceae species, such as Mru (47.54) and Msat (48.08), showed moderate ENC under low-GC3s, reflecting less extreme bias and separating compositional effects from constraint effects (Novembre, 2002).

ΔENC boxplots showed Poaceae with smaller deviations and Fabaceae with stronger negative shifts, implying gene sets with more compressed codon usage (dos Reis *et al*., 2004; Hershberg and Petrov, 2008). This suggests codon choice is influenced by selective constraints, expression levels, tRNA availability and gene functions (Fig 3).

The width of ΔENC distributions varied, with broader ranges in some species indicating highly biased gene subsets. This highlights the importance of integrating ΔENC distribution shapes to avoid misinterpreting small subsets of biased genes as genome-wide trends.

Overall, ENC-GC3s scatterplots and ΔENC boxplots reveal systematic codon-bias divergence, with structured differences between Fabaceae and non-Fabaceae lineages, providing a foundation for further optimal-codon and lineage-specific analyses (Fig 4).

PR2 analysis

PR2 analysis of 30 angiosperm species revealed significant third-position nucleotide asymmetry, with deviations from the PR2 equilibrium point (0.5, 0.5) consistent across species. The PR2 bias index (Euclidean distance to 0.5, 0.5) showed species medians above zero, indicating detectable A/T and G/C imbalances, supporting Sueoka's parity framework (Sueoka, 1995). High PR2 bias was observed in Cari, Mtru and Hvul, while species like

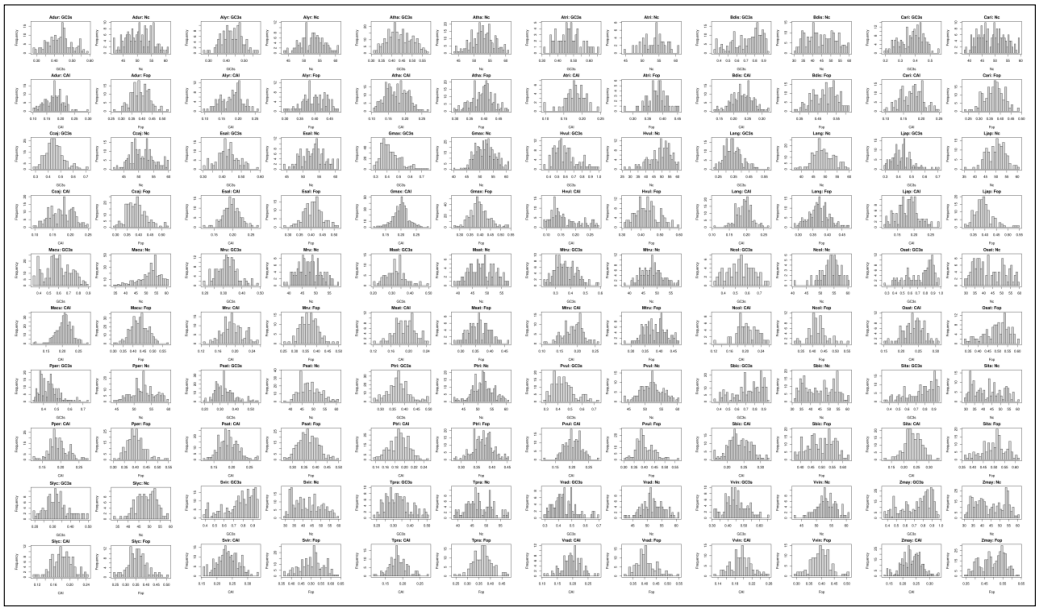


Fig 2: Gene-level distributions of codon-bias and nucleotide-composition indices across 30 angiosperm species.

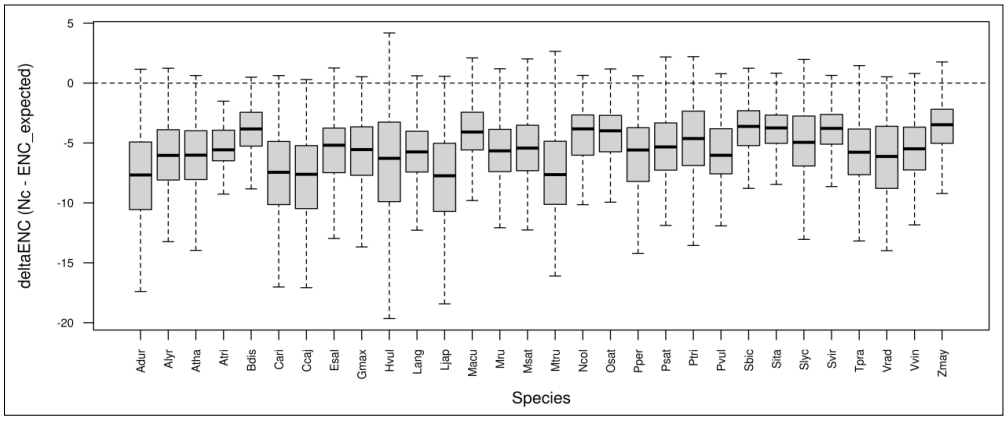


Fig 3: Cross-species deltaENC distributions quantify deviations from the Wright ENC-GC3s expectation.

Macu, Sbic and Bdis exhibited weaker deviations. This stratification suggests that third-position asymmetry is a lineage-specific compositional feature, offering a complementary axis for codon-preference differentiation independent of ENC/GC3s (Fig 5).

PR2 plots showed stable directional shifts in codon imbalances, with Fabaceae species (Adur, Cari, Ccaj) concentrated in the region of A3 enrichment and C3 depletion. In contrast, species like Atri and Ncol followed a G3/T3-enriched pattern. This indicates lineage-dependent directional biases in third-position codon usage, reflecting strand-specific mutation patterns and selective effects (Sueoka, 1995).

Within Fabaceae, PR2 bias intensity varied, with Mtru and Cari showing stronger deviations than expected based on composition alone, suggesting codon preference coupling via nucleotide frequency shifts and relative enrichment of preferred codons. This is consistent with mutation-selection-drift theory, where weak selection can drive codon usage bias (Bulmer, 1991; Hershberg and Petrov, 2008). PR2 analysis thus provides a quantitative basis for interpreting lineage differences in codon preference strength and directionality, reinforcing that codon bias cannot be explained solely by GC3 or global composition but is shaped by lineage-specific asymmetry (Fig 6).

Neutrality plot analysis

The neutrality plot (Fig 7) quantifies the relationship between GC12 (GC content at the first and second codon positions) and GC3s (GC content at synonymous third positions) to assess compositional coupling. A slope near 1 indicates mutational bias, while a slope near 0 suggests stronger constraints on the first and second positions (Sueoka, 1988; Sueoka, 1992).

Across 30 species, variation in GC3s was broader than in GC12, with Poaceae species showing wider GC3s dispersion and a positive regression trend, indicating stronger third-position shifts propagating to the first and second positions. In contrast, Fabaceae species had

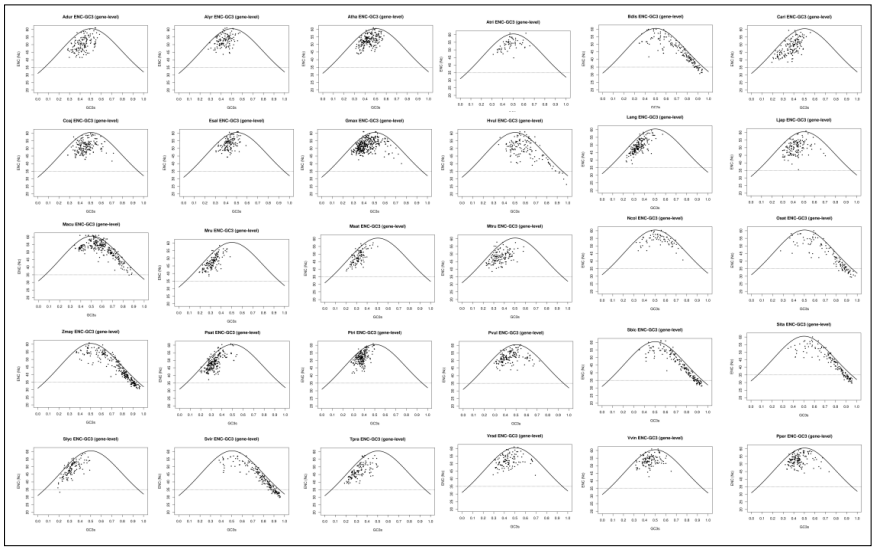


Fig 4: Species-specific ENC-GC3s patterns relative to the wright expected curve across 30 angiosperms.

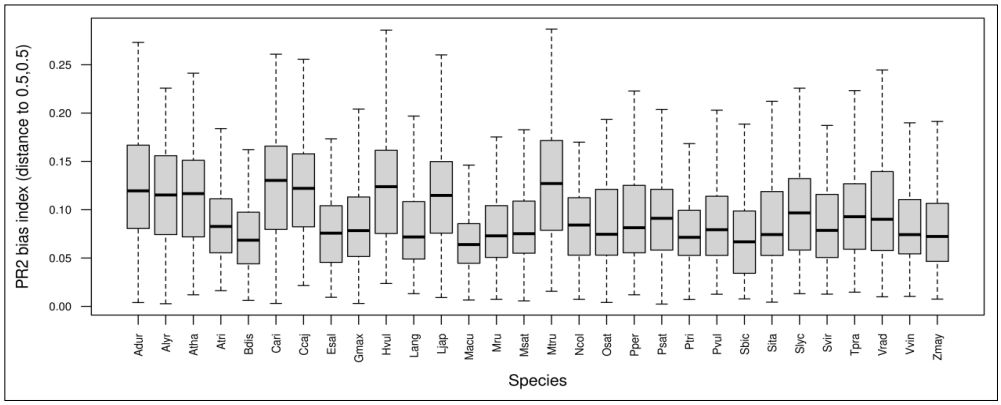


Fig 5: Species-wise distribution of PR2 bias index in CDSs across 30 angiosperm genomes.

narrower GC12 variation and weaker GC3s influence, supporting the ENC-GC3s pattern of high GC3s and low ENC in Poaceae and lower GC3s and higher ENC in Fabaceae (Plotkin and Kudla, 2011).

This partition between GC3s and GC12 suggests a "strong compositional pressure not fully neutral," with regression slopes deviating from unity. The neutrality plot thus highlights the roles of mutational bias, position-specific constraints and gene-set heterogeneity in codon-bias formation (Fig 7).

Correlation analysis comparison

At the whole-dataset level (n = 5,434 genes), GC3s was the primary axis of codon-bias variation, with the strongest correlation between Fop and GC3s (ρ = 0.754, FDR < 0.001), indicating that optimal codon usage tracks the third-position base composition gradient (Hershberg and Petrov, 2008). CAI also correlated with GC3s (ρ = 0.416, FDR < 0.001) and negatively with Nc (ρ = -0.245, FDR < 0.001), suggesting coordinated codon usage (Table 1).

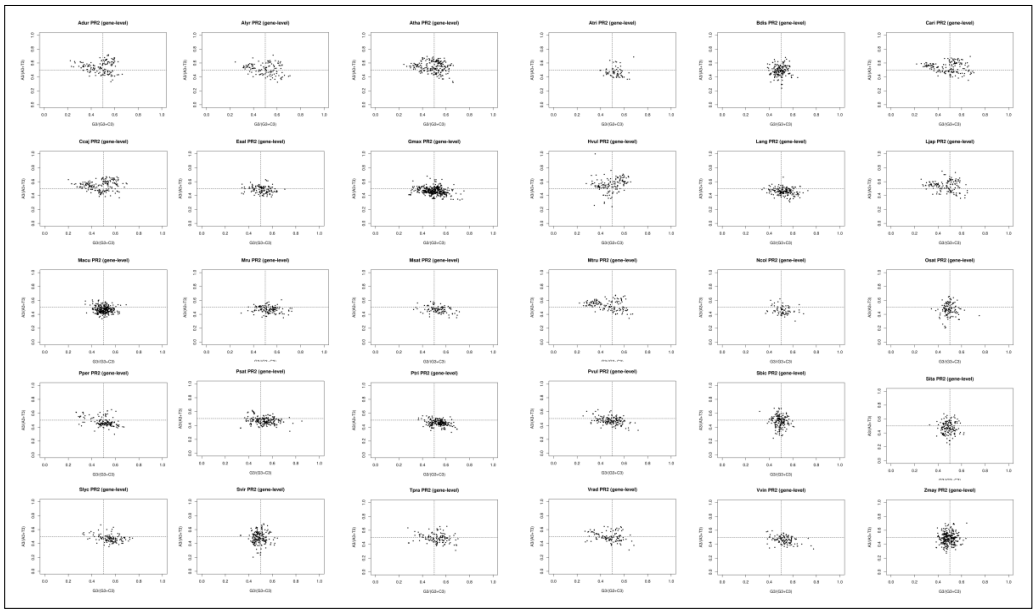


Fig 6: PR2 plots of third-codon-position base usage (A3/(A3+T3) vs G3/(G3+C3)) for 30 angiosperm species.

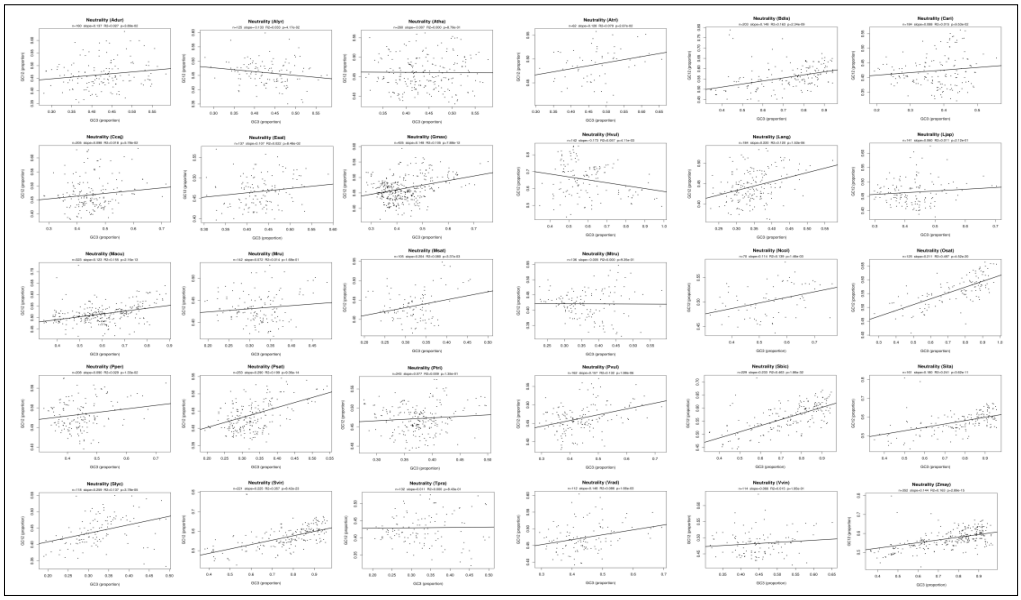


Fig 7: Species-specific neutrality plots showing the relationship between GC12 and GC3s across 30 angiosperm species.

Poaceae species (*e.g.*, Osat, Sbic) showed strong negative Nc-GC3s correlations ($\rho \approx -0.92$), while Fabaceae species exhibited more variable Nc responses. Despite a positive Fop-GC3s correlation in Fabaceae, Nc responses differed, indicating that GC3s influences optimal-codon usage without directly affecting codon concentration (Ikemura, 1981; Chamary *et al*., 2006).

DeltaENC correlated with GC3s, with higher GC3s reducing compositional deviation in some species, while reinforcing codon-bias divergence in others, suggesting additional selective constraints (Hershberg and Petrov, 2008). PR2 bias index showed lineage-specific patterns (*e.g.*, Atha, Cari, Lang).

These results highlight that GC3s is a key driver of codon-preference variation, but lineage-specific mechanisms in Nc and deltaENC require careful interpretation. The divergence between Poaceae and Fabaceae underscores the need to separate compositional influences from lineage-specific effects (Plotkin and Kudla, 2011).

RSCU clustering analysis

Hierarchical clustering of the RSCU matrix from 30 angiosperms revealed stable codon usage patterns (Fig 8). Poaceae species (*e.g.*, Osat, Svir) clustered closely, while Fabaceae species (*e.g.*, Mru, Msat) showed more variation, suggesting codon usage is influenced by lineage history and mutational/selection constraints (Bulmer, 1991; Hershberg and Petrov, 2008).

The heatmap showed coordinated shifts in codon sets, especially between G/C and A/T-ending codons, supporting compositional-driven shifts rather than local drift (Sharp and Li, 1986; Hershberg and Petrov, 2008).

PCA of RSCU profiles (Fig 9) revealed that PC1 explained 86.5% of variance, separating Poaceae from eudicots/Fabaceae. Hvul showed deviation, indicating species-specific differences. The scree plot (Fig 10) confirmed that variation is primarily captured by PC1, with PC2 and higher components reflecting lineage-specific modulation (Sharp and Li, 1987; Hershberg and Petrov, 2008).

Table 1: Correlation analysis.

Species	Nc-GC3s	CAI-Nc	CAI-GC3s	Fop-Nc	Fop-GC3s	deltaENC-GC3s	PR2_bias_index-GC3s
All	-0.094***	-0.245***	0.416***	-0.183***	0.754***	0.205***	-0.022
Adur	0.462***	0.056	-0.193*	0.345***	0.248**	0.035	0.132
Alyr	0.325***	-0.107	-0.229*	-0.053	0.082	-0.12	0.091
Atha	0.440***	-0.195**	-0.475***	0.029	0.045	0.027	0.298***
Atri	0.482***	-0.207	-0.301*	0.103	0.127	0.318*	0.085
Bdis	-0.923***	-0.438***	0.503***	-0.745***	0.811***	0.274***	0.016
Cari	0.646***	-0.091	-0.243**	0.352***	0.430***	-0.132	0.271***
Ccaj	0.284***	-0.124	-0.185*	0.111	0.241***	-0.016	-0.016
Esal	0.402***	0.085	0.18	0.211*	0.444***	-0.033	0.003
Gmax	0.431***	0.228***	0.312***	0.426***	0.566***	-0.088	0.046
Hvul	-0.424***	-0.308***	0.214*	-0.396***	0.702***	0.419***	0.031
Lang	0.714***	0.295***	0.347***	0.487***	0.438***	0.025	0.258***
Ljap	0.314***	0.035	-0.312***	0.18	0.188*	-0.11	0.099
Macu	-0.582***	-0.232***	0.282***	-0.498***	0.745***	0.205***	0.02
Mru	0.730***	0.183*	0.184*	0.311***	0.359***	-0.092	-0.124
Msat	0.665***	0.074	0.003	0.229*	0.252*	-0.027	0.012
Mtru	0.442***	-0.073	-0.190*	0.236**	0.480***	-0.484***	0.326***
Ncol	-0.213	0.082	-0.039	-0.215	0.662***	0.188	-0.109
Osat	-0.920***	-0.404***	0.472***	-0.733***	0.799***	0.146	0.266**
Pper	0.231**	-0.076	0.019	-0.022	0.468***	-0.092	0.278***
Psat	0.697***	0.391***	0.401***	0.424***	0.557***	-0.056	-0.224***
Ptri	0.457***	0.087	0.180**	0.355***	0.363***	-0.096	0.001
Pvul	0.378***	0.132	0.312***	0.273***	0.590***	-0.025	0.285***
Sbic	-0.942***	-0.611***	0.603***	-0.801***	0.833***	0.111	0.291***
Sita	-0.950***	-0.457***	0.475***	-0.728***	0.773***	0.172*	0.166
Slyc	0.720***	0.267**	0.226*	0.503***	0.549***	-0.097	0.089
Svir	-0.955***	-0.460***	0.499***	-0.737***	0.788***	0.308***	0.053
Tpra	0.710***	0.036	-0.052	0.335***	0.410***	-0.291**	0.253**
Vrad	0.334***	0.017	0.063	0.086	0.425***	-0.033	0.137
Vvin	0.288**	0.13	0.220*	0.336***	0.528***	-0.139	-0.028
Zmay	-0.932***	-0.683***	0.668***	-0.801***	0.849***	0.214***	0.269***

Correspondence analysis complements RSCU-based clustering and principal component analysis

Hierarchical clustering of the species-level RSCU matrix comprising 59 synonymous codons across 30 species revealed structured variation in codon usage (Fig 8). Species pairs such as Pper-Ccaj and Sbic-Bdis occurred in closely associated branches of the dendrogram, indicating similarity across multiple RSCU variables. Pvul and Gmax also clustered closely, whereas Atha and Alyr formed a separate association. These patterns are consistent with taxonomically structured variation in codon usage but do not, by themselves, establish phylogenetic relationships or identify the underlying evolutionary mechanisms (Grantham *et al*., 1980; Hershberg and Petrov, 2008).

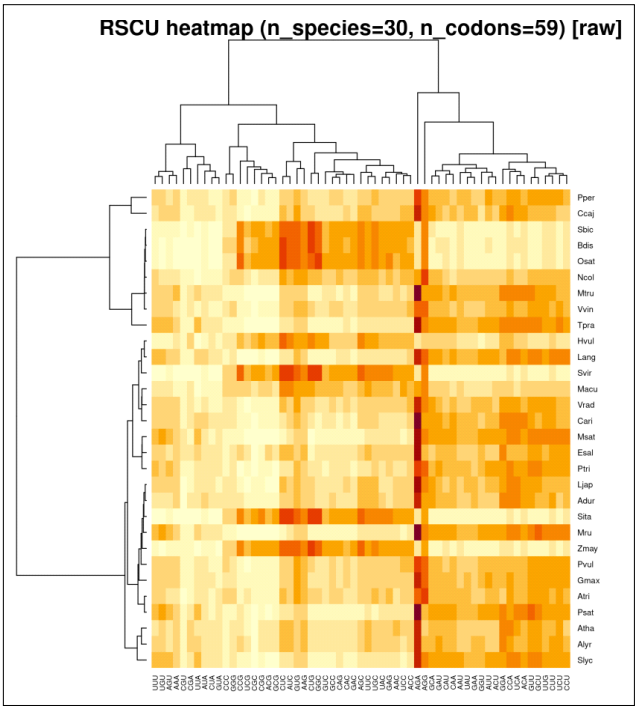


Fig 8: Species-level RSCU heatmap reveals modular codon-usage preferences and hierarchical clustering across 30 angiosperm species (n = 30, codons = 59).

PCA of the RSCU profiles showed that PC1 and PC2 explained 86.5% and 5.3% of the total variance, respectively (Fig 9). Most Poaceae species occupied the positive side of PC1, whereas most Fabaceae and other eudicot species occurred on the negative side. Macu and Ncol occupied intermediate positions along PC1, while Hvul was separated from the main Poaceae grouping primarily along PC2. The scree plot further showed that variation in the RSCU profiles was dominated by the first principal component (Fig 10). This broad separation is consistent with the established influence of genome-wide compositional processes on codon usage, although the ordination itself does not distinguish among the underlying evolutionary forces (Chen *et al*., 2004; Hershberg and Petrov, 2008).

Correspondence analysis of the species-level 64-codon count profiles provided a complementary representation of codon-count composition (Greenacre, 2007). CA1 and CA2 accounted for 75.0% and 15.6% of the total inertia, respectively, together explaining 90.6% (Fig 11 and 12). Most Poaceae species formed a compact group on the positive side of CA1, whereas Fabaceae and other eudicot species were more broadly distributed on the negative side. Hvul was clearly separated from the main Poaceae cluster along CA2. The separation among

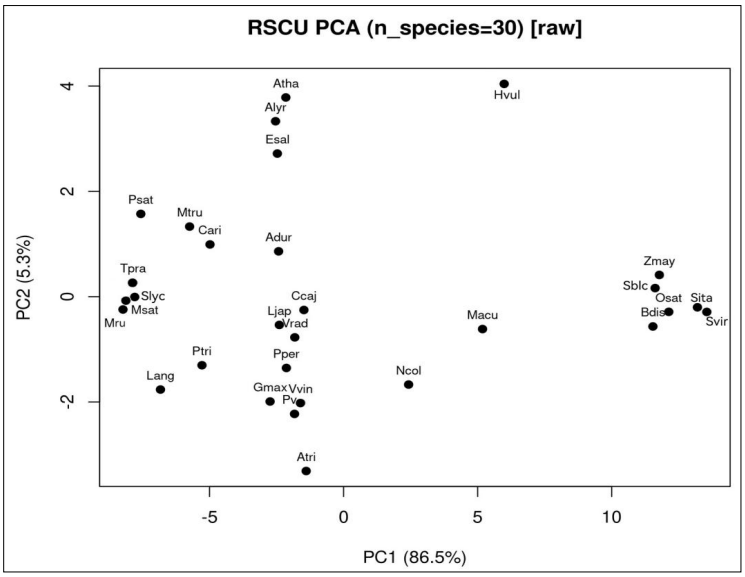


Fig 9: Principal component analysis (PCA) of species-level RSCU profiles separates angiosperm species along dominant axes of synonymous codon usage (PC1 = 86.5%, PC2 = 5.3%).

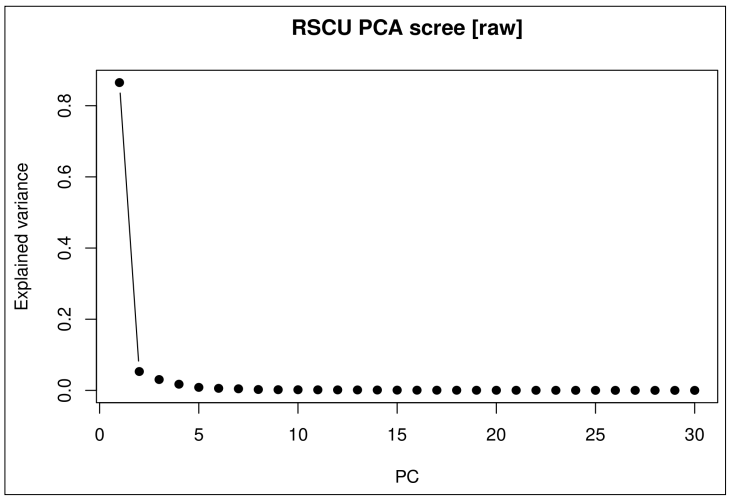


Fig 10: Scree plot of explained variance for RSCU-PCA indicates a strongly dominant first principal component.

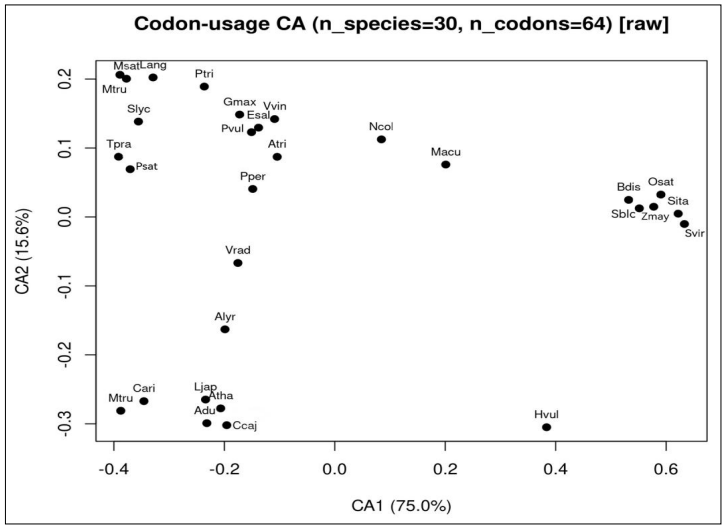


Fig 11: Two-dimensional correspondence analysis (CA) of raw codon counts across 30 angiosperm species.

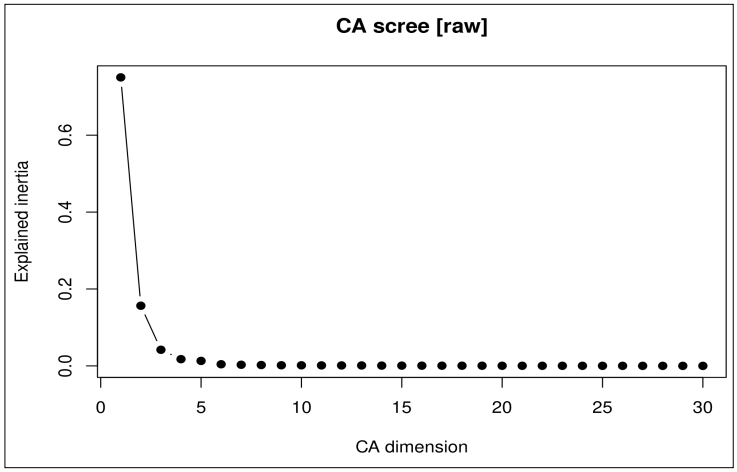


Fig 12: Scree plot of explained inertia across correspondence analysis (CA) dimensions (raw codon counts).

Fabaceae species, including Msat, Lang and Mru, further indicated substantial within-family heterogeneity in their codon-count profiles.

Taken together, hierarchical clustering, PCA and CA revealed concordant patterns of codon usage that were broadly aligned with major taxonomic groups. The agreement among these complementary analyses supports the internal consistency of the observed compositional structure across alternative representations of codon usage. However, these analyses cannot, on their own, resolve the relative contributions of genome-wide mutational bias, translational selection, gene expression, or other evolutionary processes (Ikemura, 1985; Hershberg and Petrov, 2008).

CONCLUSION

This study analyzed nuclear genome CDSs from 30 angiosperms, focusing on synonymous codon preferences in Fabaceae versus non-Fabaceae lineages. The bHLH gene family showed notable size variation (62-425 members), highlighting its role in cross-species comparisons. Poaceae species exhibited high GC3s and strong codon bias (low Nc), while Fabaceae species clustered in the low-GC3s range, reflecting compositional differences.

Codon usage bias was not solely driven by base composition. Negative deltaENC values and PR2 analysis indicated additional forces shaping codon bias. Multivariate analyses, including RSCU clustering, PCA and correspondence analysis, confirmed a consistent codon-preference structure across species.

GC3s was the dominant factor driving codon-preference divergence, with lineage-specific variations. Poaceae and Fabaceae exhibited distinct compositional and bias architectures, providing a foundation for codon optimization, expression system design and understanding synonymous site evolution.

ACKNOWLEDGEMENT

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) and the subproject “Gene Mining for Important Agronomic Traits and Germplasm Innovation in Medicago ruthenica” under the project “Research and Development of New Breeding Technologies and Promotion of New Varieties of Important Native Grasses” (Grant No. 2022JBGS0040 (0104).

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

Akashi, H. (1994). Synonymous codon usage in *Drosophila melanogaster*: Natural selection and translational accuracy. *Genetics.* **136(3):** 927-935.

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.

Bailey, P.C., Martin, C., Toledo-Ortiz, G. *et al*. (2003). Update on the basic helix-loop-helix transcription factor gene family in *Arabidopsis thaliana*. *The Plant Cell.* **15(11):** 2497-2502.

Bulmer, M. (1991). The selection-mutation-drift theory of synonymous codon usage. *Genetics*. **129(3):** 897-907.

Chamary, J.V., Parmley, J.L. and Hurst, L.D. (2006). Hearing silence: Non-neutral evolution at synonymous sites in mammals. *Nature Reviews Genetics*. **7(2):** 98-108.

Cheeran, K., Suresh, K.P., Jacob, S.S., Gowda, C.S.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.

Chen, S.L., Lee, W., Hottes, A.K., Shapiro, L. and McAdams, H.H. (2004). Codon Usage Between Genomes is Constrained by Genome-wide Mutational Processes. *Proceedings of the National Academy of Sciences of the USA.* **101(10):** 3480-3485.

dos Reis, M., Savva, R. and Wernisch, L. (2004). Solving the riddle of codon usage preferences: A test for translational selection. *Nucleic Acids Research.* **32(17):** 5036-5044.

Drummond, D.A. and Wilke, C.O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell.* **134(2):** 341-352.

Eddy, S.R. (2011). Accelerated profile HMM searches. *Plos Computational Biology*. **7(10):** e1002195.

Grantham, R., Gautier, C., Gouy, M.,Mercier, R. and Pavé, A. (1980). Codon catalog usage and the genome hypothesis. *Nucleic Acids Research*. **8(1):** R49-R62.

Greenacre, M. (2007). Correspondence Analysis in Practice. 2nd ed. Chapman and Hall/CRC, New York.

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.

Hershberg, R. and Petrov, D.A. (2008). Selection on codon bias. *Annual Review of Genetics*. **42:** 287-299.

Ikemura, T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *Journal of Molecular Biology.* **151(3):** 389-409.

Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular Biology and Evolution*. **2(1):** 13-34.

Jolliffe, I.T. (2002). Principal Component Analysis. 2nd ed. Springer, New York.

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.

Novembre, J.A. (2002). Accounting for background nucleotide composition when measuring codon usage bias. *Molecular Biology and Evolution*. **19(8):** 1390-1394.

Peden, J.F. (1999). Analysis of Codon Usage. Ph.D Thesis, University of Nottingham, Nottingham, UK.

Peng, X., Mu, Y., Wu, F., Fu, N., Shi, F. and Zhang, Y. (2025). Analysis of codon preferences in *Medicago ruthenica* based on transcriptome data. *Legume Research.* **48(7):** 1145-1153. doi: 10.18805/LRF-776.

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. *et al.* (2015). Codon optimality is a major determinant of mRNA stability. *Cell.* **160(6):** 1111-1124.

Schmutz, J., Cannon, S.B., Schlueter, J. *et al*. (2010). Genome sequence of the palaeopolyploid soybean. *Nature*. **463(7278):** 178-183.

Schnable, P.S., Ware, D., Fulton, R.S. *et al*. (2009). The B73 maize genome: Complexity, diversity and dynamics. *Science.* **326(5956):** 1112-1115.

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

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.

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

Sueoka, N. (1992). Directional mutation pressure, selective constraints and genetic equilibria. *Journal of Molecular Evolution.* **34(2):** 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.

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