Integrative Genomic and Machine Learning Approaches Reveal Evolutionary Signatures in the Winged Bean Mitochondrial Genome

Nikhil Kumar Singh
ICAR-Indian Institute of Agricultural Biotechnology
Binay K. Singh
ICAR-Indian Institute of Agricultural Biotechnology
Piyush Kumar
ICAR-Indian Institute of Agricultural Biotechnology
Avinash Pandey
ICAR-Indian Institute of Agricultural Biotechnology
Sudhir Kumar
ICAR-Indian Institute of Agricultural Biotechnology
Sujit Kumar Bishi
ICAR-Indian Institute of Agricultural Biotechnology
A. Pattanayak
ICAR-Indian Institute of Agricultural Biotechnology
V. P. Bhadana
ICAR-Indian Institute of Agricultural Biotechnology
Sujay Rakshit
ICAR-Indian Institute of Agricultural Biotechnology
Kishor U. Tribhuvan
kishor.tribhuvan@icar.gov.in

ICAR-Indian Institute of Agricultural Biotechnology

Research Article

Keywords: Winged bean, mitochondrial genome, synteny, codon bias, phylogenetics, sustainable agriculture, machine learning, organelle genome classification

Posted Date: October 3rd, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7571989/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Plant Biology on November 28th, 2025. See the published version at https://doi.org/10.1186/s12870-025-07738-6.

Abstract

The mitochondrial genome (mitogenome) of *Psophocarpus tetragonolobus* (winged bean), a nutritionally valuable yet genomically underexplored tropical legume, was assembled using high-coverage PacBio long reads and Illumina short reads. The 366,925 bp circular genome encodes 64 genes (38 protein-coding, 20 tRNAs, 6 rRNAs) and contains nine fragmented protein-coding genes, indicative of dynamic mitogenome architecture. Repeat profiling revealed 100 dispersed repeats (30–110 bp) and 25 SSRs (4.95% of the genome), with assembly graph inspection and recombination models supporting subgenomic circles and isoforms. Comparative analyses across 15 legumes showed pervasive purifying selection, with positive selection in specific codons of *atp4*, *ccmB*, *cox1*, *nad3*, and *rps10*. Codon usage bias differed markedly between organelles: mitochondrial genes exhibited moderate bias consistent with neutral expectations, whereas chloroplast genes showed greater variability, suggesting additional selective constraints. Synteny mapping revealed multiple conserved and inverted regions between organelles, highlighting structural divergence. Leveraging 14 codon bias metrics, we implemented the first machine learning framework for organelle genome classification in plants, achieving up to 0.96 AUC and identifying GC3s as the most influential feature. This integrative genomic, evolutionary, and ML-based approach advances understanding of *P. tetragonolobus* mitogenome evolution and establishes a proof-of-concept with potential cross-kingdom applications.

Introduction

Plant mitochondrial genomes are essential to cell function, powering energy production, coordinating cellular signaling, and supporting responses to environmental stress [1]. Their genetic system is semi-autonomous [1, 2] and markedly different from that of animals, showing greater structural complexity, frequent recombination, and episodes of horizontal gene transfer [3, 4]. These genomes are often large and variable in size due to the accumulation of non-coding sequences, gene duplications, and the integration of foreign DNA from nuclear or chloroplast origins [5]. Although such complexity poses challenges for assembly and annotation, it offers valuable insights into the mechanisms shaping plant genome evolution and driving structural rearrangements [6]. Much of this variation arises from dynamic processes that actively remodel the mitochondrial genome over time.

Structural diversity in plant mitochondrial genomes is largely driven by recombination across repeated sequences, generating multipartite configurations in circular or linear forms [3]. These repeats drive homologous recombination, generating subgenomic molecules that vary in copy number and contribute to genome plasticity [7]. Such flexibility helps maintain genome integrity under stress and may enhance adaptation by increasing genetic diversity within the organelle population [8]. Additional complexity arises from introns, which influence gene regulation and splicing, particularly under fluctuating environmental conditions [9]. Horizontal gene transfer (HGT) from nuclear and chloroplast genomes can further expand the functional repertoire of mitochondria, introducing genes that improve resilience to both biotic and abiotic stressors [5, 10]. Together, these processes generate unique structural and functional signatures that can be compared across species to trace evolutionary histories and adaptive

trajectories. In addition to structural variation, organelle genomes also differ in their coding sequence composition, with measurable biases in synonymous codon usage that reflect evolutionary pressures and functional constraints. These sequence-level features provide complementary information to structural patterns and can be quantitatively analysed to uncover organelle-specific signatures.

Codon usage bias is a powerful genomic signature shaped by mutational pressures, translational selection, and tRNA availability, and it carries evolutionary and functional information at both the gene and genome level. These biases differ systematically between organelles such as mitochondria and chloroplasts, offering a measurable “compositional fingerprint” that can be exploited for accurate prediction of gene origin using advanced machine learning (ML) algorithms [11]. We hypothesised that ML could leverage codon usage features to robustly discriminate between organelle genomes, providing a novel framework for organelle gene annotation, evolutionary analysis, and detection of horizontal gene transfers. To test this hypothesis, we focused on the winged bean (*Psophocarpus tetragonolobus*), an underutilised yet nutritionally dense legume native to Southeast Asia, Africa, and the Pacific Islands. This climate-resilient crop is valued for its unique edibility—leaves, flowers, seeds, and tubers are all consumed—and its seeds contain up to 35% protein, rivalling soybean, while the tubers are rich in starch and essential minerals [12, 13]. Its adaptability to tropical environments, coupled with high nutritional content and ecological benefits from nitrogen fixation, make it a promising candidate for food security and climate-resilient agriculture.

Lastly, the winged bean’s mitochondrial genome is of evolutionary interest. Plant mitochondrial genomes are highly variable in size and structure across species, with significant rearrangements, repeated sequences, and foreign DNA integrations from nuclear and chloroplast genomes [14, 15]. Studying the winged bean’s mitochondrial genome can provide insights into the evolutionary relationships within the legume family and identify genomic adaptations unique to this species. Such research could illuminate the evolutionary pressures and genetic changes that have enabled winged bean to thrive in tropical conditions, contributing to the broader understanding of plant mitochondrial evolution and adaptation [16, 17].

In this study, we report the complete assembly and annotation of the *P. tetragonolobus* mitochondrial genome and perform a comprehensive comparative analysis with related legumes. We characterize its structural organization, repeat landscape, phylogenetic relationships, and patterns of selection, alongside an in-depth evaluation of codon usage bias in both mitochondrial and chloroplast genes. Building on these compositional features, we apply benchmark multiple supervised machine learning algorithms to assess whether codon usage metrics alone can accurately classify organelle gene origin. This represents, to our knowledge, the first proof-of-concept application of ML to organelle genome classification in plants, integrating genome architecture, evolutionary analyses, and computational prediction. Our findings provide novel insights into organelle genome evolution and deliver a valuable genomic resource to advance crop improvement and the breeding of climate-resilient, nutritionally rich legume varieties.

Results

Mitochondrial genome assembly and gene content

We assembled the *Psophocarpus tetragonolobus* mitochondrial genome into a single 366,925-bp sequence with ~ 871× mean coverage, supporting high base-level confidence (Fig. 1; Supplementary Fig. S1). For this we used filtered high-quality Illumina short reads (523303993 reads) and PacBio long reads (3709181 reads, 63.7 Gbp). The assembly and raw reads are available under BioProject PRJNA1255599, and the mitogenome has been deposited in GenBank as PQ877725.1. The overall GC content is 45.44%, consistent with angiosperm mitogenomes.

The mitochondrial genome contains 64 genes, including 38 protein-coding genes, 20 tRNA genes, and 6 rRNA genes (Fig. 1, Table 1). Among the protein-coding genes, key groups include ATP synthase subunits, NADH dehydrogenase subunits, cytochrome genes, and ribosomal protein genes (Fig. 1). The mitochondrial genome contains 38 protein-coding genes, which are categorized into distinct functional groups (Table 1). The ATP synthase subunits are represented by 5 genes (*atp1, atp4, atp6, atp8, and atp9*), encoding components of the ATP synthase complex that drive ATP production through oxidative phosphorylation. The NADH dehydrogenase subunits are encoded by 11 genes (*nad1, nad2, nad3, nad4, nad4L, nad5, nad6, nad7, nad9*), which are essential components of Complex I in the electron transport chain, playing a key role in transferring electrons and establishing the proton gradient for ATP synthesis. The cytochrome genes include 4 genes (*cox1, cox2, cox3, and cob*), which encode subunits of cytochrome c oxidase (Complex IV) and cytochrome b (Complex III), critical for electron transfer and maintaining the mitochondrial energy cycle. Additionally, the genome includes ribosomal protein genes, with 7 genes (*rps3, rps4, rps12, rpl2, rpl5, rpl16, and rpl10*) encoding structural components of mitochondrial ribosomes required for efficient protein synthesis within mitochondria. Interestingly, nine fragmented protein-coding genes are presented in the mitogenome, reflecting the dynamic nature of mitochondrial genome organization. These fragmented genes include *nad4, nad7, nad2* (with a fragmented form, *nad2*-fragment), *nad5* (with nad5-fragment), *nad1* (with nad1-fragment), *atp6* (with fragmented forms *atp6*), *rps4-fragment, rps10 and rps3*. The 20 tRNA genes ensure efficient mitochondrial translation, while the 6 rRNA genes, likely reflecting a plant mitochondrial origin, may include fragmented or duplicated forms typical of plant mitochondria (Table 1).

Table 1

a: Annotated genes in mitochondrial genome of *P. tetragonolobus*

Sr. No	Gene	Strand	Start	End
Protein coding Genes				
1	*atp1*	Reverse	326051	327577
2	*atp4*	Forward	202783	203370
3	*atp6*	Reverse	68826	69535
4	*atp8*	Reverse	324315	324360
5	*atp9*	Reverse	338892	339116
6	*ccmB*	Reverse	321766	322392
7	*ccmC*	Reverse	25767	26504
8	*ccmFc**	Reverse	246236	251634
9	*ccmFn*	Reverse	76031	77767
10	*cob*	Reverse	175192	176363
11	*cox1*	Reverse	157372	158955
12	*cox3*	Reverse	163341	164138
13	*matR*	Reverse	304915	306920
14	*mttB*	Reverse	322620	323342
15	*nad1**	Forward	49794	260262
16	*nad2**	Reverse	281230	286203
17	*nad3*	Reverse	189396	189752
18	*nad4**	Reverse	109451	118076
19	*nad4L-1*	Forward	202280	202582
20	*nad5**	Forward	191530	193839
21	*nad6*	Forward	53526	54192
22	*nad7**	Forward	86915	91122
23	*nad9*	Forward	71609	72181
24	*rp15*	Reverse	177604	178161
25	*rpl16*	Forward	263336	263851
*Genes having introns				

Sr. No	Gene	Strand	Start	End
26	*rps1*	Forward	190715	191332
27	*rps10**	Forward	204545	207722
28	*rps12*	Reverse	188970	189347
29	*rps14*	Reverse	177294	177596
30	*rps3**	Forward	260377	263445
31	*rps4*	Forward	69536	70492
rRNA Genes				
32	*rrn5*	Forward	235791	235911
33	*rrnL*	Forward	33137	34819
34	*rrnS*	Forward	233620	234990
tRNA Genes				
35	trnQ-UUG | trnQ-TTG	11699	11770	Reverse
36	trnI-CAT	15732	15813	Forward
37	trnM-CAU	15732	15808	Forward
38	trnS-GGA	19740	19818	Forward
39	trnM-CAU	36310	36383	Forward
40	trnnull-NNN	54135	54194	Forward
41	trnH-GUG	81949	82022	Reverse
42	trnTERM-UUA	104194	104263	Forward
43	trnD-GUC	120065	120138	Forward
44	trnP-UGG	139538	139612	Reverse
45	trnF-GAA	139873	139946	Reverse
46	trnS-GCT	140302	140388	Reverse
47	trnS-GCU	140303	140388	Reverse
48	trnA-UGC*	211849	211953	Reverse
49	trnV-GAC	214931	215002	Reverse
50	trnW-CCA	231777	231850	Forward
*Genes having introns				

Sr. No	Gene	Strand	Start	End
51	trnM-CAU	301038	301110	Reverse
52	trnE-UUC	302003	302074	Reverse
53	trnK-UUU	308715	308787	Reverse
54	trnC-GCA	319726	319796	Forward
*Genes having introns				

Table 1
b: List of fragmented copies of genes in mitochondrial genome of *P. tetragonolobus*

Sr. No	Gene	Strand	Start	End
Protein coding Genes				
1	*nad1-fragment*	Reverse	304169	307616
2	*nad7-fragment*	Reverse	257316	257578
3	*rps4-fragment*	Reverse	162131	162324
4	*rps4-fragment*	Forward	189877	189993
5	*atp6-1-fragment*#	Reverse	26732	26839
6	*atp6-1-fragment*#	Reverse	321556	321664
7	*atp9-fragment*	Forward	97090	97139
rRNA Genes				
8	*rrnS-fragment*	Forward	234991	235621
9	*rrnL-fragment*	Forward	32786	33136
10	*rrnL-fragment*	Forward	34820	35933
#Gene fragments present in two copies				

Repetitive Sequences and Large Repeat Regions

We identified 100 large repeats in the *P. tetragonolobus* mitogenome (40 forward, 60 palindromic) ranging from 30–110 bp, with both longest forward and palindromic repeats measuring 108 bp (Fig. 2a-c, Supplementary Table S2). The overall mitochondrial genome has a GC content of 45.44%, while the average GC content of the repeat regions is slightly higher at 47.41% (Supplementary Table S2). These repeats span 18,157 bp, representing 4.95% of the 366,925 bp genome, and have a higher GC content (47.41%) than the genomic average (45.44%). Of these, 51 overlap coding regions including *atp6*, *nad1*,

and fragmented loci and display higher GC content than intergenic repeats (n = 49), which are more variable in composition (Fig. 2d; Supplementary Fig. S2). We also detected 25 simple sequence repeats (SSRs) and 19 variable number tandem repeats (VNTRs) (Fig. 2e–f). SSR motifs comprised monomers (12%), dimers (36%), and trimers (52%), while VNTRs (≥ 85% identity) ranged from 10–39 bp in length.

Bandage visualisation of the assembly graph revealed looped and bifurcating paths, indicative of repeat-mediated homologous recombination (Supplementary Fig. S3). These features are consistent with alternative genome conformations, such as subgenomic circles and inversion isoforms, and with multipartite mitochondrial architectures previously reported in angiosperms. Schematic models illustrate the potential rearrangements mediated by large direct and palindromic repeats (Supplementary Fig. S4), providing evidence for a structurally dynamic *P. tetragonolobus* mitogenome capable of generating multiple isoforms through recombination.

Codon Usage Bias Analysis

The codon usage analysis of the mitochondrial genome of *P. tetragonolobus* revealed significant codon bias, with distinct preferences for specific codons and amino acids (Fig. 3). Among the most frequently used codons were TTT (Phenylalanine) and ATT (Isoleucine), with counts of 392 and 300 and RSCU values of 1.15 and 1.347, respectively, highlighting their preferential usage in mitochondrial protein synthesis (Fig. 3, [18, 19]). Conversely, codons such as CTG (Leucine) and GAC (Aspartic acid) were underutilized, with low RSCU values of 0.535 and 0.573. Leucine, encoded by six codons, showed a strong bias for CTT (RSCU = 1.429) and TTA (RSCU = 1.234), while CTG was significantly underrepresented. Similarly, serine exhibited a preference for TCT (RSCU = 1.32) and TCA (RSCU = 1.116), indicating selective use of these codons. Stop codons displayed a clear hierarchy of usage, with TGA (RSCU = 1.126) being the most frequent, followed by TAA (RSCU = 1.025) and TAG (RSCU = 0.849). Codons with RSCU values > 1, such as GAT (Aspartic acid, RSCU = 1.427) and GAA (Glutamic acid, RSCU = 1.289), were overrepresented, whereas those with values < 1, such as GCG (Alanine, RSCU = 0.637), were underutilized, reflecting translational optimization (Fig. 3). Overall, the codon usage bias reflects mitochondrial-specific selective pressures, favoring translational efficiency and compatibility with mitochondrial tRNA pools, likely optimizing protein synthesis and reducing translational errors.

Phylogenetic Analysis

The phylogenetic analysis of the mitochondrial genome of *P. tetragonolobus* reveals its evolutionary relationships within the Fabaceae family and its comparison with tuber crops (Fig. 4). Tuber crops were included in the analysis due to the winged bean's similarity to both legumes and tuber crops. The tree demonstrates clear clustering within the Fabaceae family, with species such as *Medicago truncatula*, *Trifolium* spp., *Lotus japonicus*, *Phaseolus vulgaris*, *Vigna* spp., and *Glycine* spp. forming well-supported clades, as indicated by bootstrap values ranging from 0.89 to 1.0 (Fig. 4). *P. tetragonolobus* clusters closely with legumes such as *G. max* (bootstrap value: 0.95) and *V. radiata* (0.93), underscoring their

shared lineage within the Phaseoleae tribe. Its relationship with *G. soja* is similarly well-supported (bootstrap value: 0.87), reflecting genetic similarity among Phaseoleae members. Further divergence is observed within Fabaceae, with *P. tetragonolobus* grouping more distantly with *P. vulgaris* (bootstrap value: 0.82) and *V. angularis* (0.79). Beyond Phaseoleae, the Trifolieae clade comprising *T. pratense*, *Trifolium aureum*, and *T. grandiflorum* shows strong separation (bootstrap values 0.85–0.91), indicating differentiation at the tribal level. *M. truncatula* and *A. hypogaea* also cluster within Fabaceae but are more distantly related to *P. tetragonolobus*. Other legumes such as *S. flavescen*s, *S. tora*, *T. indica*, and *B variegata* occupy intermediate evolutionary positions, consistent with subfamily-level divergence. Outside the Fabaceae family, *P. tetragonolobus* exhibits clear divergence from tuber crops, including *S. tuberosum* (bootstrap value: 0.61) and *I. batatas* (bootstrap value: 0.58). These tuber crops form distinct clusters, highlighting their evolutionary distance from legumes. *M. esculenta*, although a tuber crop, is more distantly placed.

Synteny analysis

Synteny analysis using pairwise BLAST results were conducted to evaluate genomic relationships and conservation among *P. tetragonolobus, G. max, V. radiata, P. vulgaris, M. truncatula* (Fig. 5). These analyses revealed key conserved and variable regions, providing insights into evolutionary dynamics (Fig. 5). The mitochondrial genome of *G. max* (401,300 bp) showed 92% syntenic coverage with *P. tetragonolobus*, characterized by conserved blocks ranging from 10 kb to 150 kb (Supplementary table S2). *V. radiata* (379,425 bp) and *P. vulgaris* (377,980 bp) shared 88% and 86% syntenic coverage, respectively, with conserved blocks spanning 20 kb to 120 kb. Notably, synteny between *P. vulgaris* and *V. radiata* included large, uninterrupted conserved blocks (Fig. 5). This level of conservation reflects a shared evolutionary history and strong functional constraints on mitochondrial gene order (Fig. 5) and Phylogeny (Fig. 4). In contrast, *M. truncatula* (317,280 bp) exhibited fragmented synteny, with only 65% of the *P. tetragonolobus* genome covered by smaller conserved blocks of 5 kb to 50 kb. Percent identity distributions revealed that across all datasets, alignments predominantly fell within the 90% to 98% range, indicating moderate to high conservation. The *P. tetragonolobus vs. G. max* dataset peaked sharply at 98% with also the lowest E-value, while *G. max vs. V. radiata* had a similar peak at 97.8% with E-value below − 150 (Supplementary Fig S3).

We compared 34 core mitochondrial genes across eight legume species, including *P. tetragonolobus*. The heatmap (Supplementary Fig. S5) revealed high conservation in key respiratory (*atp, cox, nad, cob*) and cytochrome maturation (*ccm*) genes, with *P. tetragonolobus* retaining a complete set. Ribosomal protein genes (*rps, rpl*) showed greater variability, with occasional absences such as rps14, likely reflecting lineage-specific gene loss or nuclear transfer. Hierarchical clustering based on gene presence and absence grouped *P. tetragonolobus* with *G. max* and *V. radiata*, consistent with its placement within the Phaseolinae clade. Strand orientation analysis of 79 genes identified 22 (~ 28%) with inversion events across species (Supplementary Table. S3), predominantly in ribosomal and RNA-associated genes (e.g., *rps12, rpl5, rrnL, matR*). In contrast, all major respiratory and maturation genes maintained

consistent orientation in *P. tetragonolobus*, reinforcing the structural integrity of its mitochondrial genome. These findings suggest that while peripheral components of the mitogenome are prone to variation, core functional regions remain conserved across legumes.

Selection pressure analysis

Initial pairwise dN/dS (Ka/Ks) estimates from KaKs_Calculator3 revealed that most *P. tetragonolobus* mitochondrial genes are under strong purifying selection, particularly in comparisons with closely related legumes such as *G. max* and *G. soja* (e.g., *atp1*, Ka/Ks = 0.0249), indicating strong functional constraints (Fig. 6a). Comparisons with more distantly related species, including *Arabidopsis thaliana* and *Raphanus sativus*, yielded moderately higher ratios (0.40–0.41), reflecting relaxed purifying selection. Certain genes, such as *atp4*, displayed Ka/Ks ratios approaching neutrality (0.91–0.96) with *Vigna radiata* and *Brassica rapa*, while others like *ccmB* exhibited values > 1 in specific comparisons (e.g., 2.27 with *Manihot esculenta*), suggesting lineage-specific adaptive divergence.

To incorporate phylogenetic context, we applied CodeML, which confirmed that purifying selection predominates across the mitogenome (ω < 1 in most branches; Fig. 6b), but identified elevated ω values in *atp4*, *ccmB*, *rps10*, *matR*, and *nad9*, consistent with episodic relaxation or adaptive change. Site-model analyses (M8: Beta + ω > 1) pinpointed specific codons under diversifying selection (Table S2). In *atp4*, codons 141 (K), 144 (T), 151 (A), and 153 (S) showed strong BEB support (P ≥ 0.998) with mean ω > 8.9, while positively selected sites were also detected in *ccmB*, *cox1*, *nad3*, and *rps10*. These codons occur in genes integral to oxidative phosphorylation and cytochrome function, suggesting possible adaptation in response to species-specific metabolic or cytonuclear interactions. Together, these results indicate that while purifying selection maintains core mitochondrial gene functions, specific residues in key energy-related proteins have undergone adaptive change, particularly in lineages more distantly related to *P. tetragonolobus*, underscoring a balance between functional conservation and evolutionary innovation.

Comparative analysis of winged bean Chloroplast and mitochondrial genomes

The relationship between the effective number of codons (ENC; range: 20 = extreme bias to 61 = no bias) and GC content at the third codon position (GC3s) was examined for *P. tetragonolobus* mitochondrial and chloroplast genes (Fig. 7a). Mitochondrial genes (ENC = 40–60; GC3s = 0.25–0.75) clustered closely to the expected ENC curve, indicating moderate codon bias primarily shaped by mutational pressures. In contrast, chloroplast genes (ENC = 30–55; GC3s = 0.00–0.75) showed a wider scatter from the curve, suggesting that translational selection and functional constraints additionally influence codon usage.

Parity Rule 2 (PR2) analysis (Fig. 7b) further supported these trends. Mitochondrial genes were tightly clustered near the equilibrium point (A3/[A3 + T3] = 0.45–0.55; G3/[G3 + C3] = 0.40–0.60), consistent with mutational bias dominance and balanced nucleotide composition at third codon positions. Conversely, chloroplast genes spanned a broader range (A3/[A3 + T3] = 0.30–0.70; G3/[G3 + C3] = 0.25–

0.75), with many deviating from equilibrium, indicating substantial influence from selective pressures related to photosynthetic and other essential functions.

A circular synteny plot (Fig. 7c) highlighted homologous regions, gene annotations, and inversion synteny between the mitochondrial and chloroplast genomes. In mitochondria, syntenic loci included *trnE-TTC* (302,003–302,074; 78 bp; 94.87% identity), *trnM* (301,038–301,110; 78 bp; 94.87%), *trnW* (231,777–231,850; 98 bp; 95.91%), and *trnD* (120,065–120,138; 82 bp; 98.78%). In the chloroplast genome, conserved synteny corresponded to *trnH*, *trnP*, and inverted repeat regions IR-A and IR-B, with inversions affecting loci such as *ndhB* and *rrnL*. Percent identities for syntenic alignments generally exceeded 90%, with alignment lengths ranging from ~ 80 bp in functional tRNA regions to > 300 bp in large, conserved blocks. Visualization integrates gene names, alignment statistics, and inversion patterns, providing a high-resolution view of conserved and rearranged organellar architecture.

Codon Usage Patterns Predict Organelle Origin with High Accuracy through Machine Learning

Principal component analysis (PCA; Fig. 8a) of 14 codon usage–related features revealed a clear separation between P. tetragonolobus chloroplast and mitochondrial genes, with PC1 (37.4% variance) driven by GC3s, total GC content, and third-position base composition, and PC2 (16.6% variance) influenced by Nc, aromaticity, and GRAVY. This unsupervised clustering confirmed that codon bias metrics alone capture biologically meaningful distinctions between organelle types. The same separation was preserved under non-linear dimensionality reduction using UMAP (Fig. 8b), indicating that these compositional differences are robust to both linear and non-linear transformations.

Prediction probability density plots (Fig. 8c) showed minimal overlap between chloroplast and mitochondrial classes in probability space, reflecting high model confidence in its predictions prior to formal accuracy assessment. To quantify classification performance, six supervised algorithms: rf, svmRadial, gbm, xgbTree, nnet, and nb were benchmarked (Table 3, See Methods for details), yielding mean ROC AUC values of 0.89–0.96, mean accuracies of 0.87–0.93, and kappa values of 0.65–0.80. The xgbTree model was selected as optimal because it achieved the highest mean ROC AUC (0.960) while maintaining balanced sensitivity (0.943) and specificity (0.683), which is essential when class imbalance could otherwise skew performance.

On the *P. tetragonolobus* training dataset (n = 144; 108 chloroplast, 36 mitochondrial; Supplementary Table S4), the xgbTree model achieved perfect classification (accuracy = 1.0, kappa = 1.0, AUC = 1.0; Fig. 8d), demonstrating complete separability of the two organelle types within this species. Independent validation on 367 genes from diverse plant species (246 chloroplast, 121 mitochondrial, Supplementary Table S5)) resulted in an accuracy of 83.4%, sensitivity of 86.6%, specificity of 76.9%, and a balanced accuracy of 81.7% (Fig. 8e). Misclassifications comprised 28 mitochondrial genes predicted as chloroplast and 33 chloroplast genes predicted as mitochondrial, reflecting reduced generalization when applied to phylogenetically distant taxa.

Variable importance analysis (Fig. 8f) identified GC3s as the most influential predictor (relative importance = 100.0), followed by total GC content (59.8), G3s (56.7), and A3s (42.5), underscoring the central role of codon position–specific GC bias in differentiating organelle genomes. Nc (38.6) and aromaticity (26.4) contributed substantially, while GRAVY (23.1) and L_sym (18.9) had moderate influence. L_aa (0.1) contributed negligibly, indicating that protein length offers little discriminatory power once compositional biases are considered.

Discussion

The successful assembly and annotation of the mitochondrial genome of the winged bean (*P. tetragonolobus*) mark a significant advancement in the genomic understanding of this underutilised crop. With its total size of 366,925 bp and inclusion of 64 genes, comprising 38 protein-coding genes, 12 tRNA genes, and 6 rRNA genes, the winged bean mitochondrial genome reveals several features characteristic of angiosperm mitochondrial genomes. These findings provide valuable insights into the molecular underpinnings of energy metabolism, stress tolerance, and evolutionary dynamics in this nutritionally and agriculturally important legume.

Gene Content Conservation in Plant Mitochondria

The variability in mitochondrial genome size among plants reflects a delicate balance between maintaining essential functions and adapting to environmental challenges. Despite the variance in mitogenome sizes, ranging from approximately 66 kb [20] to over 11 Mb [21] across plant species, the functional genes in mitochondrial genomes are highly conserved [22]. These conserved genes primarily encode components of the electron transport chain, ATP synthesis, and mitochondrial translation machinery, ensuring the core energy-producing functions of mitochondria remain intact [23, 24]. In *P. tetragonolobus*, protein-coding genes are categorized into key groups: ATP synthase subunits, NADH dehydrogenase subunits, cytochrome-related genes, and ribosomal protein genes. These genes encode enzymes and structural proteins vital for oxidative phosphorylation and ATP production, highlighting the functional specialization of mitochondria in energy metabolism and cellular homeostasis.

Consistent with this conserved functional core, the mitochondrial genome of *P. tetragonolobus* includes a complete set of protein-coding genes, classified into ATP synthase subunits, NADH dehydrogenase subunits, cytochrome-related genes, and ribosomal protein genes. These genes encode key enzymes and structural components for oxidative phosphorylation and cellular energy metabolism. The gene inventory closely mirrors that of *Glycine max* and *Vigna radiata*, indicating deep conservation of mitochondrial respiratory and translational capacity within Phaseolinae legumes. Minor variations in ribosomal protein genes, such as rps14 or rpl16, are consistent with relaxed selective pressure or postulated nuclear transfer events frequently observed in legume mitogenomes.

Structural synteny analysis further revealed extensive collinearity between P. tetragonolobus and G. max, supported by high-identity BLASTn alignments (> 98%) across multi-kilobase regions. In contrast, more

distantly related legumes such as *P. vulgaris* and *M. truncatula* exhibited substantial structural rearrangements, including inversions and block translocations, particularly within intergenic and repeat-rich regions. These findings are consistent with the high structural plasticity characteristic of plant mitogenomes and suggest that such rearrangements are tolerated as long as coding sequences remain intact.

The structural and functional integrity of the *P. tetragonolobus* mitochondrial genome may underlie its notable resilience in tropical environments. Although direct links between mitogenome architecture and metabolic performance are not yet established, the retention of all core genes, along with conserved genome structure, suggests a stable and robust organellar system [25, 26]. Future studies incorporating transcriptomic and physiological data could illuminate the role of mitochondrial genome architecture in stress adaptation and ecological fitness.

Beyond *P. tetragonolobus*, mitochondrial genomics has broad implications for legume breeding and stress adaptation. Mitochondrial genome analysis has yielded crucial insights into both abiotic stress responses [27] and cytoplasmic male sterility (CMS) [28] in legumes, traits that can be leveraged for crop improvement. For example, in soybean, comparative sequencing of CMS lines' mitochondria identified novel ORFs (e.g., orf178 and orf261) associated with pollen sterility [29]. Similarly, in pigeonpea (*C. cajan*), comparative mitochondrial genomics uncovered multiple chimeric ORFs linked to CMS, enabling a unique CMS-based hybrid breeding system in this crop with major yield potential [28]. Beyond reproductive traits, organellar genomics also illuminates abiotic stress resilience: chickpea (*C. arietinum*) under salinity stress shows a "stress-modified" mitochondrial electron transport chain, including coordinated upregulation of alternative oxidase and NADH dehydrogenase genes – which correlates with enhanced salt tolerance [27]. Likewise, in the model *M. truncatula*, stress conditions induce the accumulation of specific mitochondrial protein complexes (such as prohibitins) that mediate abiotic stress tolerance and trigger retrograde signaling [30]. These examples highlight the applied significance of mitochondrial genome insights, spanning improved stress resilience and developing CMS-based hybrids, for legume breeding programs

Structural Complexity of the Winged Bean Mitochondrial Genome

The sizeable mitochondrial genome sizes observed in plants are often attributed to extensive repetitive sequences, which drive homologous recombination [31, 32]. In *P. tetragonolobus*, 100 repeat sequences, including 40 forward and 60 palindromic repeats, underscore the potential recombination's role and thus help shape the mitochondrial genome's structure. Homologous recombination events generate subgenomic molecules and facilitate structural rearrangements, which are both adaptive and functionally significant [31]. Similar observations in legumes such as *P. vulgaris* and *M. truncatula* suggest conserved mechanisms of mitochondrial genome evolution within the Fabaceae family [31, 33–35]. Additionally, identifying simple sequence repeats (SSRs) and variable number tandem repeats (VNTRs) in the winged bean's mitochondrial genome highlights its structural diversity. These repetitive

elements enhance genomic plasticity and serve as valuable molecular markers for population genetics and evolutionary studies [36, 37].

Comparable to other legumes, the mitochondrial genome of P. tetragonolobus exhibits distinct features indicative of its adaptability to challenging environments. Studies in related species, such as Glycine max and Vigna radiata, have reported similarly dynamic mitochondrial architectures, reflecting lineage-specific responses to environmental pressures [21, 32, 38–40]. Notably, the higher repeat density observed in *P. tetragonolobus* suggests an elevated potential for recombination-driven genomic plasticity, which may contribute to its exceptional resilience under tropical stress conditions. This observation is consistent with findings in Lotus japonicus, where repetitive elements correlate with enhanced genome flexibility and stress adaptation [15, 41].

The presence of numerous large forward and palindromic repeats, along with the complex topologies observed in the assembly graph (Supplementary Fig. S3), supports the hypothesis of repeat-mediated conformational rearrangements. Such conformational plasticity is a hallmark of many plant mitochondrial genomes. Multipartite or structurally dynamic mitogenomes have been documented in species including Vigna radiata, Cucumis sativus, and Silene noctiflora, where recombination across large repeats produces subgenomic circles, inversions, and stoichiometric shifts [21, 40]. In Brassica species, extensive mitochondrial rearrangements have been functionally linked to altered gene expression and cytoplasmic male sterility, underscoring the biological impact of structural variants.

Our hypothetical models (Supplementary Fig. S4) further illustrate the recombination potential of the *P. tetragonolobus* mitogenome. This structural flexibility likely contributes to fine-tuned mitochondrial function under varying physiological conditions and may facilitate rapid adaptation in response to environmental stressors. While further experimental validation (e.g., Southern blotting or long-read phasing) is needed, our data strongly support the presence of a structurally dynamic mitochondrial genome with evolutionary parallels across legumes and other angiosperms.

Codon usage bias and tRNA evolution in the mitogenome of P. tetragonolobus

The codon usage analysis of the mitochondrial genome of *P. tetragonolobus* reveals significant preferences for specific codons, correlating with its tRNA availability. This reflects evolutionary optimization for efficient mitochondrial protein synthesis [42–44]. Among the most frequently used codons are TTT (Phenylalanine) and ATT (Isoleucine), with RSCU values of 1.15 and 1.347, respectively. These biases towards A/T-ending codons are consistent with patterns observed in other plant mitochondrial genomes, such as those of *G. max* and *A. thaliana* [45, 46] and have been previously observed in the chloroplast genome of *P. tetragonolobus* [47]. Such preferences are linked to mutational biases and selective pressures favouring translational efficiency. Additionally, the presence of tRNA genes such as trnK-TTT, trnE-TTC, and trnM in *P. tetragonolobus* aligns with the observed codon usage patterns. For example, trnK-TTT supports the preferential usage of TTT codons for phenylalanine, enhancing translational accuracy and efficiency. Similarly, underrepresented codons like CTG (Leucine, RSCU = 0.535) and GAC (Aspartic acid, RSCU = 0.573) may reflect limited availability or functionality of

the corresponding tRNAs, as seen in other legumes where translational constraints shape codon bias [4]. In *A. thaliana*, codon usage biases are similarly influenced by tRNA availability, but distinct codon preferences, such as a higher reliance on G/C-ending codons, highlight species-specific adaptations [48]. In contrast, *G. max* shows broader codon usage diversity, reflecting its more complex tRNA repertoire [48].

Stop codon usage in *P. tetragonolobus* shows a clear hierarchy, with TGA (RSCU = 1.126) being the most frequent, followed by TAA (RSCU = 1.025) and TAG (RSCU = 0.849). This pattern aligns with findings in other plants, where stop codons are optimized for termination efficiency and error minimization [21, 49]. The preference for TGA is particularly notable and may reflect its compatibility with mitochondrial release factors. In summary, the codon usage bias and tRNA gene content of *P. tetragonolobus* highlight the interplay between evolutionary constraints and translational optimization. These results, consistent with other plant mitochondrial genomes trends, provide insights into evolutionary dynamics shaping mitochondrial functionality in legumes and other plants.

Inferring evolutionary dynamics of the winged bean mitochondrial genome

The mitochondrial genome of *P. tetragonolobus* provides valuable insights into its evolutionary trajectory within the Fabaceae family, while also offering a first glimpse into its divergence from tuber crops despite morphological similarities. Phylogenetic analysis reveals robust clustering of *P. tetragonolobus* within the Fabaceae family, closely associating it with *G. max* (bootstrap value: 0.97) and *V. radiata* (0.93) in the Phaseoleae tribe. Its relationship with *G. soja* (0.86) underscores shared ancestry within Phaseoleae and highlights genetic diversity between domesticated (*G. max*) and wild (*G. soja*) soybean species. Divergence from more distantly related Fabaceae, such as *M. truncatula* (0.69) and *L. japonicus* (0.79), reflects differentiation at the tribal level, supported by whole-genome and chloroplast genome phylogenies [47, 50]. Despite morphological similarities, *P. tetragonolobus* is evolutionarily distinct from tuber crops like *S. tuberosum* (0.58) and *I. batatas* (0.52), which form separate clusters indicative of independent evolutionary paths.

Phylogenetic and synteny analyses together provide a comprehensive view of evolutionary dynamics within plant species [51–53]. Our synteny analysis complements phylogenetic findings, revealing high genomic conservation among *P. tetragonolobus* and closely related legumes. The mitochondrial genome of *G. max* shares 92% syntenic coverage with *P. tetragonolobus*, with conserved blocks spanning 10 kb to 150 kb. Similarly, *V. radiata* (88%) and *P. vulgaris* (86%) exhibit substantial synteny, reflecting strong functional constraints and shared evolutionary history. In contrast, *M. truncatula* displays fragmented synteny, with 65% coverage and smaller conserved blocks (5 kb to 50 kb), signifying more significant divergence. These structural patterns parallel our gene content and orientation findings: *P. tetragonolobus* retains all core respiratory and maturation genes with consistent strand orientation, reflecting a stable and functionally robust mitogenome. Conversely, strand inversions in ribosomal and RNA-related genes across species point to lineage-specific rearrangements in less constrained regions.

Together, these results underscore the dual nature of plant mitochondrial genome evolution conserved in core functional domains but flexible in non-essential regions and support the evolutionary resilience of *P. tetragonolobus* in the Phaseolinae clade.

The Ka/Ks ratio analysis of mitochondrial genes provides critical functional insights, reinforcing patterns observed in synteny and phylogeny. Genes like *atp1* exhibit consistently low Ka/Ks ratios (0.0249) in comparisons with closely related species such as *G. max* and *G. soja*, reflecting strong purifying selection that supports phylogenetic clustering and syntenic conservation in our study. Higher Ka/Ks ratios observed with distantly related species, such as *A. thaliana* (0.4100) and *M. esculenta* (2.2739), suggest relaxed purifying selection or adaptive divergence, consistent with fragmented synteny in these lineages. Similarly, moderate Ka/Ks ratios in comparisons with *V. radiata* (0.3007) and *P. vulgaris* (0.3328) indicate functional constraints in line with their shared evolutionary history. Positive selection signals, such as those observed in *ccmB* with *M. esculenta*, align with functional divergence and phylogenetic distance, whereas conserved genes like *ccmC* with low Ka/Ks ratios emphasize strong evolutionary constraints that maintain essential mitochondrial functions. These findings align with previous studies using organelle genomes in species like Mangifera [54], Brassica [54], *Punica granatum L* [55], *Cucumis sativus* [56] and other legumes [40, 47, 57–61]

This variation could point to evolutionary processes like domestication and natural selection shaping these lineages [62, 63]. *P. tetragonolobus* diverged into the Phaseoleae tribe alongside *G. max* (soybean) and *V. radiata* (mung bean) [47, 50] adapting to specific environments with distinct traits—climbing habits in winged beans versus bush growth in soybeans. A similar divergence is observed in grasses like *Oryza sativa* (rice) and *Zea mays* (maize), where rice adapted to water-logged tropical conditions and maize evolved for temperate dryland farming [64–66]. These findings collectively emphasize the evolutionary importance of maintaining mitochondrial gene functionality across lineages. The combined phylogenetic, synteny, and Ka/Ks analyses contribute to understanding mitochondrial genome evolution and its role in shaping lineage-specific adaptations. Future studies should emphasize functional analyses to explore the adaptive significance of conserved and divergent genomic regions.

Evolutionary insights from codon usage and genomic architecture of organelle genomes

The codon usage and synteny analyses of mitochondrial and chloroplast genomes in *P. tetragonolobus* reveal complementary evolutionary dynamics driven by distinct pressures and functional roles. ENC-GC3s analysis shows mitochondrial genes clustering near the expected ENC curve, with moderate codon bias (ENC 40–60, GC3s 0.25–0.75) predominantly shaped by mutational pressures [67–69]. In contrast, chloroplast genes display greater variability (ENC 30–55, GC3s 0.00–0.75), suggesting additional influences like translational selection and functional constraints related to photosynthesis [70–73]. PR2 analysis supports these findings: mitochondrial genes cluster near equilibrium (G3/(G3 + C3) and A3/(A3 + T3) around 0.5), reflecting mutational dominance. Chloroplast genes exhibit broader distributions (G3/(G3 + C3) 0.25–0.75, A3/(A3 + T3) 0.3–0.7), indicating a balance of mutational biases

and selective pressures [46, 74–76]. In contrast, *A. thaliana* displays a broader range of GC3s values and stronger codon bias for certain genes, reflecting more pronounced selective pressures and functional constraints in its mitochondrial genome [49].

Our machine learning analysis demonstrates that codon usage bias alone can robustly distinguish chloroplast from mitochondrial genes, achieving high cross-validation performance (ROC AUC 0.89–0.96; accuracy 0.87–0.93) and 83.4% accuracy on a multi-species test set, with GC3s emerging as the most informative feature (relative importance = 100), followed by total GC content (59.8), G3s (56.7), and A3s (42.5). This alings with the large-scale ML trained on ~ 13,000 organisms has already shown near-perfect discrimination of organelle origin (nuclear vs. mitochondrial vs. chloroplast) and high taxonomic accuracy from codon frequencies, with PCA confirming that most variance tracks A/T- vs. G/C-ending codons precisely the pattern our feature rankings captured via GC3s and related third-position metrics [77]. Cross-kingdom studies similarly report that codon usage alone classifies yeasts to order with > 90% accuracy, indicating that lineage-specific mutation/selection regimes leave reproducible, machine-learnable signatures in codon composition [78]. Mechanistically, these signals arise from the well-established balance of mutation pressure and translational selection shaping synonymous sites (including tRNA pool matching, translational speed/accuracy, and mRNA structure effects), with genome-wide GC (and specifically GC3s) a dominant axis of variation [18, 79]. In plants, chloroplast and mitochondrial coding sequences systematically differ in codon usage and third-position composition, where chloroplast genes typically prefer A/U-ending codons under stronger selective constraints related to photosynthesis, whereas plant mitochondria exhibit distinct, often more muted biases providing a priori biological separability that ML can exploit [80]. Recent organelle-pair analyses (e.g., sweet potato) explicitly demonstrate cp vs. mt divergence in RSCU and base composition, consistent with our model's reliance on GC3s/GC as top discriminants [80]. Together, these findings support a practical hypothesis: codon composition encodes a conserved "compartmental signature" reflecting mutation-bias spectra and translational selection tuned to organelle-specific biochemistry and tRNA ecology; ML leverages this signature to (i) annotate ambiguous organelle genes in new assemblies, (ii) flag potential horizontal gene transfers that violate host codon context, and (iii) quantify lineage- or environment-specific shifts in translational regimes across taxa [77].

The circular synteny plot highlights high conservation in mitochondrial genes such as *trnE-TTC* and *trnM* (> 94% identity), aligning with their balanced codon usage. Chloroplast regions like IR-A and IR-B show inversion synteny and structural rearrangements, reflecting their adaptive variability. Percent identities above 90% across homologous regions demonstrate shared evolutionary origins, while alignment lengths (~ 80 to 300 bp) underscore lineage-specific adaptations. Mitochondrial genomes exhibit stability driven by mutational pressures and functional constraints, supporting energy metabolism. Chloroplast genomes show more significant variability, shaped by selection pressures linked to photosynthesis and environmental adaptation. The interplay between codon usage and structural conservation underscores the evolutionary strategies balancing conservation and adaptability in organellar genomes. Future studies should integrate functional data to explore these patterns further [54, 81–85].

The mitochondrial genome of *P. tetragonolobus* offers critical insights into the evolutionary dynamics and functional adaptations of organellar genomes. Codon usage patterns highlight the role of mutational pressures in shaping mitochondrial genes, while chloroplast genes display greater variability driven by translational selection and functional constraints. Synteny analysis underscores high genomic conservation, interspersed with lineage-specific structural rearrangements that reflect evolutionary divergence. These findings collectively emphasize the interplay of conservation and adaptability in organellar genomes, shedding light on the molecular mechanisms underlying stress tolerance, energy metabolism, and environmental resilience in this underutilized crop. Future research integrating functional, transcriptomic, and proteomic data will further elucidate the adaptive significance of these evolutionary patterns and support the sustainable utilization of *P. tetragonolobus* in agriculture.

Methods

Plant material, Sampling and DNA extraction

The cultivar AKWB1 (IC0508386) was procured from the All India Coordinated Research Project on Potential Crops, ICAR-National Bureau of Plant Genetic Resources (NBPGR), New Delhi, India. To ensure genetic uniformity, AKWB1 was self-pollinated for three generations prior to the start of the experiment. Plants were grown under field conditions at the ICAR–Indian Institute of Agricultural Biotechnology (IIAB), Ranchi (23°16'27.6"N, 85°20'29.4"E). Leaf samples from 25-day-old plants were harvested for genome sequencing. High-quality total genomic DNA was extracted using the cetyltrimethylammonium bromide (CTAB) method [86]. To remove RNA contamination 2.0 µl of RNase A (10 mg/ml, HiMedia) was added to 20 µl of DNA dissolved in TE buffer (Tris–EDTA, pH = 8.0), followed by incubation at 37°C for 3–4 h. The whole-genome sequencing library was constructed using the TruSeq DNA PCR-Free Library Prep Kit following the manufacturer's instructions. After assessing the quality and quantity of the library, whole-genome sequencing was conducted on Illumina HiSeq2500 instrument with paired-end (2 × 150 bp) sequencing strategy using the TruSeq Rapid SBS Kit. The complete mitochondrial genome sequence and raw Illumina reads have been deposited in the NCBI under BioProject accession number PRJNA1255599. Parallely, the same biomass was also used to perform long read PacBio sequencing. Briefly, the library was constructed using SMRTbell Express template Preparation Kit 2.0 and purified with AMPure PB beads per the manufacturer's protocol. About 80 pM of the library was loaded onto one SMRTcell containing 8M ZMW and sequenced in PacBio Sequel II system in CCS/HiFi mode.

Sequencing quality assessment

Illumina short reads were checked for quality using FastQC v. 0.12.0 [87] and read counts were extracted. Sequencing reads were then trimmed for adapter sequences and sequencing quality using Trimmomatic v. 0.39 [88] using the following settings: illuminaclip = TruSeq3-PE.fa:2:30:10, leading = 10, trailing = 10, sliding-window = 5:10 and minlen = 50. The raw subreads obtained from the PacBio Sequel II sequencer were processed to generate highly accurate circular consensus sequences (HiFi reads) using the CCS (Circular Consensus Sequencing) tool [89].

Mitochondria genome assembly and annotation

We assembled the mitogenome using trimmed and filtered PacBio long reads with the MitoHiFi tool v.3.2.2 [90]. We choose *Glycine max* (PRJNA927338) as the closest reference as it was shown to be the closest species when comparing the chloroplast genome [91]. The model organism was set to plant with "*-a plant*" option. This step yielded multiple mitochondrial contigs; the longest complete circular configuration was selected as a seed reference. In the second step, we used this seed to reassemble the mitochondrial genome with GetOrganelle tool v.1.7.7.1 [92] and short reads specifying the embryophyta plastid database and using the candidate mito genome fasta file. The parameters used were "*-k 21,45,65,85,105 -P 1000000 -w 85*". The Maximum extension rounds were set to 20. The resultant assembly in the graphical fragment assembly (GFA) was manually checked in Bandage [93]. Contigs with coverage of less than 50 or those that mapped to the chloroplast genome were removed, and the retained graph was exported in fasta format. which allowed iterative extension and polishing using all available HiFi reads. This approach combined the high accuracy of MitoHiFi's reference-guided contig generation with GetOrganelle's graph-based recovery of complete organellar genomes. The final assembly was validated by read mapping, coverage analysis, and gene content assessment.

Annotation of the mitochondrial genome was then conducted, referencing all previously published mitochondrial genomes of Fabaceae family, utilising Geseq software for this purpose [94]. All GeSeq annotations were validated by sequence similarity and domain evidence. Protein-coding genes were confirmed by TBLASTN/BLASTX against RefSeq plant mitochondrial proteins (E-value ≤ 1e − 10; alignment length ≥ 70% of query; identity ≥ 35%).The output genebank and gff files were manually checked, corrected, and a circular representation was drawn using OGDRAW [95]. The complete mitochondrial genome sequence of the winged bean (*Psophocarpus tetragonolobus* L.) was submitted to GenBank with the accession number PQ877725.1.

Investigation of DNA repeat sequences

For the identification of repetitive sequences within the genome, microsatellites, tandem repeats, and dispersed repeats were conducted using Krait v1.3.3 software [96], Tandem Repeats Finder [97] and REPuter [98], respectively. Repeats were reported as F (forward) or P (palindromic) at a Hamming Distance of 3 and a Minimal Repeat Size of 30. The chloroplast genome was also screened for perfect SSRs (pSSRs), compound SSRs (cSSRs), and variable number tandem repeats (VNTRs) using Krait v1.3.3 software [96]. The screening was conducted based on specific criteria: mono-nucleotide repeat motifs were required to have a minimum of 10 repeats, di-nucleotide repeat motifs required a minimum of five repeats, tri-nucleotide repeat motifs needed at least five repeats, while tetra-, penta-, and hexa-nucleotide repeat motifs were required to have a minimum of four repeats. If the distance between two SSRs was < 10 bp, they were considered as cSSR.

Phylogeny and synteny analysis

Along with annotated mitochindrial genome of *P. tetragonolobus* a total of 27 plant species, primarily from the Fabaceae family, were selected for phylogenetic analysis based on conserved mitochondrial gene sequences. Complete or nearly complete mitochondrial genomes were retrieved from GenBank for the following species: *Ammopiptanthus mongolicus* (NC_039660.1), *Ammopiptanthus nanus* (NC_046466.1), *Arachis hypogaea* (MW448460.1), *Astragalus membranaceus* (OR567492.1), *Bauhinia variegata* (OR188023.1), *Glycine max* (NC_020455.1), *Glycine soja* (NC_039768.1), *Lotus corniculatus* (PP706441.1), *Lotus japonicus* (NC_016743.2), *Medicago truncatula* (NC_029641.1), *Phaseolus vulgaris* (NC_045135.1), *Senna tora* (NC_038053.1), *Sophora flavescens* (NC_043897.1), *Styphnolobium japonicum* (PP035769.1), *Tamarindus indica* (NC_045038.1), *Trifolium aureum* (NC_048502.1), *Trifolium grandiflorum* (NC_048501.1), *Trifolium pratense* (NC_048499.1), *Vicia faba* (KC189947.1), *Vigna angularis* (NC_021092.1), and *Vigna radiata* (NC_015121.1). In addition, representative reference species from other plant families were included to root the phylogeny and enhance resolution: *Arabidopsis thaliana* (NC_037304.1), *Brassica rapa* (NC_049892.1), *Ipomoea batatas* (NC_068714.1), *Manihot esculenta* (NC_045136.1), and *Solanum tuberosum* (NC_059127.1). These accessions were used to extract conserved protein-coding mitochondrial genes for multiple sequence alignment and Maximum Likelihood phylogenetic reconstruction under the GTR + G model with 1000 bootstrap replicates. The sequences were aligned using MAFFT v7.487 [99, 100] with 1000 iterative refinement steps using "--maxiterate 1000". The resulting aligned sequences were saved in FASTA format. Phylogenetic trees for each aligned gene were inferred using RAxML-NG v.1.2.2 [101] with 1000 Bootstrap replicates and the "General Time Reversible (GTR) with gamma-distributed rate variation among sites (GTR + G)" model. The best gene tree files were used to prepare a multigene-based species tree with ASTRAL v5.7.8 63 [102, 103]. Finally, the phylogenetic tree was visualised using FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

We then used the mitogenomes of *G. max*, *V. radiata, P. vulgaris, M. truncatula*, and analyzed the synteny and rearrangements in the mitogenomes of *P. tetragonolobus.* For synteny plots, pairwise BLASTN results were generated using the whole mitogenomes of each species in R using a custom script. Information on BLAST hits was visualised in R using the genoPlotR package [104]. Information on BLAST hits among homologous segments was also visualised in R using the genoplotR package [104]. BLAST results were also summarised in R.

Estimation of adaptive evolution of protein-coding genes

To analyze the degree of non-synonymous (Ka) and synonymous (Ks) substitutions, as well as their ratio (dN/dS), the coding DNA sequences (CDS) of *P. tetragonolobus* were compared with 16 other species including *Glycine max*, *Glycine soja*, *Vigna radiata*, *Phaseolus vulgaris*, *Lotus japonicus*, *Medicago truncatula*, *Trifolium aureum*, *Trifolium grandiflorum*, *Trifolium pratense*, *Brassica rapa*, *Manihot esculenta*, *Ipomoea batatas*, *Raphanus sativus*, *Solanum tuberosum.*, along with *A. thaliana*. For this, we performed pairwise alignments using MAFFT v7 [99], and then Ka, Ks substitutions and their ratio were calculated using KaKs_Calculator 3 [105] with the MA model.

Codon-based evolutionary analyses were performed using the codeml module of PAML v4.9 [106, 107]. Codon alignments for mitochondrial genes of Psophocarpus tetragonolobus were generated using MAFFT, followed by manual trimming to remove poorly aligned regions. For pairwise Ka/Ks estimation, codeml was run in one-ratio mode (runmode = -2, model = 0) with F3x4 codon frequencies. To detect positively selected codons, we applied the site model M8 (NSsites = 8) using gene-wise phylogenetic trees inferred via RAxML-NG under the GTR + G model. Codons with BEB posterior probabilities ≥ 0.95 were considered under positive selection. All results were parsed and visualized using custom R and Python scripts.

Codon usage analysis

The codon usage within the coding part of the chloroplast was calculated using CodonW v1.4.4 available through https://codonw.sourceforge.net, using universal codon standards. These included T3, C3, A3, and G3, representing nucleotide usage frequencies at the codons' third base. Additionally, GC3s refer to the GC content in the third position of synonymous codons, and the ENc (effective number of codons) was also calculated using CodonW. The GC content of the first (GC1), second (GC2) and third codons(GC3) was determined using the online CUSP tool available at https://www.bioinformatics.nl/cgi-bin/emboss/cusp.

The parity rule 2 (PR2) plot was utilized to explore the nucleotide composition at the third position of codons. This commonly adopted method determines whether mutation or selection pressure dominates the nucleotide composition in DNA double strands. In the PR2 plot, the horizontal and vertical coordinates represent G3/(G3 + C3) and A3/(A3 + T3), respectively. If only mutation pressure influences the codons, all the points are in the center of the plot. Alternatively, if codon usage is influenced by natural selection and other factors, it causes the points to deviate from the center of the plot.

In the ENc-GC3s plot, the GC3s is plotted on the horizontal axis, while the actual ENc values (actENc) are plotted on the vertical axis. Notably, codons encoding methionine (Met) and tryptophan (Trp) were excluded due to their absence of synonymous codons. The expected ENc (expENc) values were obtained using the following formula:

$$expENc = 2 + GC3s + 29/(GC3 + (1 - GC3s)^2)$$

A standard curve was generated based on the expected values. Data points on the standard curve indicate a significant influence of mutational pressure on codon usage. Conversely, data points significantly below the standard curve indicate that natural selection and other factors may be the primary driving forces influencing codon usage patterns (ref). ENC ratio (ENc ratio) was determined using the following formula, which illustrates the difference between the actENc and expENc values:

ENcratio=(expENc − actENc)/expENc

Neutrality plot analysis was conducted to investigate the extent of influence between mutation pressure and natural selection on codon usage patterns. GC12 (the average GC content at the first and second

positions of codon), GC1, GC2, and GC3 were analyzed. The GC3 content was plotted on the horizontal axis, while the GC12 content was plotted on the vertical axis, resulting in two-dimensional scatter plots representing individual genes. In the neutrality plot, if the slope of the regression curve approaches 0, and there is no significant correlation between GC12 and GC3, it shows that natural selection is the primary factor driving codon usage patterns. Conversely, when the slope is close to or equal to 1, and a significant correlation exists, it indicates that mutation pressure is likely to exert a significant influence on gene evolution.

Machine Learning–Based Prediction of Organelle Gene Origin from Codon Usage Bias

Dataset Preparation and Dimensionality Reduction

Protein-coding genes from the *P. tetragonolobus* chloroplast (n = 108) and mitochondrial (n = 36) genomes were extracted from our curated organelle genome annotations (Supplementary Table S4). For external validation, we compiled a test dataset of 367 organelle genes from diverse angiosperm species (246 chloroplast, 121 mitochondrial) obtained from NCBI RefSeq, ensuring representation across monocots and dicots to capture phylogenetic diversity. Only complete coding sequences without internal stop codons were retained. Principal component analysis (PCA) and uniform manifold approximation and projection (UMAP) were applied to visualize feature space structure. PCA was performed using the prcomp function in R on z-score–scaled features, while UMAP used the umap package with parameters n_neighbors = 15 and min_dist = 0.1. These techniques were chosen to capture both linear (PCA) and nonlinear (UMAP) separations between chloroplast and mitochondrial genes, allowing assessment of whether codon bias alone encoded discriminative information.

Model Training and Evaluation

We benchmarked six supervised classification algorithms—Random Forest (rf), radial kernel Support Vector Machine (svmRadial), Gradient Boosting Machine (gbm), Extreme Gradient Boosting Tree (xgbTree), Artificial Neural Network (nnet), and Naive Bayes (nb)—using the caret v6.0 framework in R [108]. Models were trained on the *P. tetragonolobus* dataset (n = 144) with 10-fold cross-validation repeated three times to obtain stable performance estimates. Hyperparameters were tuned via grid search optimizing ROC AUC. Performance metrics included mean ROC AUC, accuracy, Cohen's kappa, sensitivity, and specificity. The xgbTree model was selected as optimal based on the highest ROC AUC while maintaining balanced sensitivity and specificity, which is critical for class-imbalanced datasets [109]. Model interpretability was assessed via variable importance ranking using xgb.importance from the xgboost v1.6 package [110, 111].

The final xgbTree model was retrained on the complete *P. tetragonolobus* dataset and tested on the 367-gene external dataset to evaluate generalization across phylogenetic distances. Misclassification patterns were quantified as false positives (mitochondrial genes predicted as chloroplast) and false

negatives (chloroplast predicted as mitochondrial). All statistical computations were conducted in R v4.3.1. ROC curves were generated with the pROC v1.18 package [85]. Balanced accuracy was calculated as the mean of sensitivity and specificity. All figures were produced using ggplot2 v3.4.4 [112].

Chloroplast to mitochondrion DNA transformation

DNA migration is common in plants and varies from species to species. This phenomenon occurs during autophagy, gametogenesis, and fertilization. The previously published chloroplast genome of the winged bean [91] was compared with the mitochondrial sequence using Blastn software [113] and visualised using Circoletto [114] to identify the synteny between the organelle genomes Screening criteria were set as the matching rate ≥ 70%, E-value ≤ $1e^{-10}$, and length ≥ 40.

Abbreviations

ICAR
Indian Council of Agricultural Research
NBPGR
National Bureau of Plant Genetic Resources
IIAB
Indian Institute of Agricultural Biotechnology
CTAB
Cetyltrimethylammonium Bromide
RNase A
Ribonuclease A
TE buffer
Tris–EDTA Buffer
ZMW
Zero–Mode Waveguide
GFA
Graphical Fragment Assembly
SSRs
Simple Sequence Repeats
VNTRs
Variable Number Tandem Repeats
PAML
Phylogenetic Analysis by Maximum Likelihood
CDS
Coding DNA Sequence
ENC
Effective Number of Codons

PCA
Principal Component Analysis
UMAP
Uniform Manifold Approximation and Projection
ORF
Open Reading Frame.

Declarations

Ethics approval and consent to participate

This study was conducted following all ethical guidelines and principles and was approved by the competent authority of the Institute. All participants provided informed consent before participating in the study.

Consent of publication

Not applicable

Availability of Data and Materials

The raw sequencing data and mitochondrial genome have been submitted to the NCBI under BioProject accession number PRJNA1255599 and PQ877725.1.

Competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

No funds have been sanctioned for this work.

Authors' contributions

NKS-data analysis and writing manuscript, BKS- sequencing, and writing the manuscript, PK-data analysis, AP-data analysis, SK-Sample generation and sequencing, SKB-data analysis, APt- Coordinated the project, manuscript editing, VBP- Coordinated the project, manuscript editing, SR- Coordinated the project, manuscript editing, KUT-Planning the experiment, sequencing, and writing the manuscript

Acknowledgements

The authors express their immense gratitude to Nucleome informatics pvt. Ltd for their invaluable support and collaboration in this research. Their expertise, resources, and technical assistance greatly contributed to the success of this study.

References

1. Douce R. Mitochondria in Higher Plants: Structure, Function, and Biogenesis. Elsevier; 2012.
2. Day D, Harvey Millar A, Whelan J. Plant Mitochondria: From Genome to Function. Springer Science & Business Media; 2013.
3. Bock R, Knoop V. Genomics of Chloroplasts and Mitochondria. Springer Science & Business Media; 2012.
4. Mower JP, Sloan DB, Alverson AJ. Plant Mitochondrial Genome Diversity: The Genomics Revolution. Plant Genome Diversity Volume 1. Vienna: Springer Vienna; 2012. pp. 123–44.
5. Gualberto JM, Newton KJ. Plant Mitochondrial Genomes: Dynamics and Mechanisms of Mutation. Annu Rev Plant Biol. 2017;68:225–52.
6. Kubo T, Newton KJ. Angiosperm mitochondrial genomes and mutations. Mitochondrion. 2008;8:5–14.
7. Maréchal A, Brisson N. Recombination and the maintenance of plant organelle genome stability. New Phytol. 2010;186:299–317.
8. Arrieta-Montiel MP, Mackenzie SA. Plant Mitochondrial Genomes and Recombination. Plant Mitochondria. 2011;:65–82.
9. Daniell H, Chase CD. Molecular Biology and Biotechnology of Plant Organelles: Chloroplasts and Mitochondria. Springer Science & Business Media; 2007.
10. Adams KL, Palmer JD. Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. Mol Phylogenet Evol. 2003;29:380–95.
11. Khomtchouk BB, Wahlestedt C, Nonner W. A global perspective of codon usage. bioRxiv. 2016.
12. Rakvong T, Monkham T, Sanitchon J, Chankaew S. Tuber development and tuber yield potential of winged bean (Psophocarpus tetragonolobus (L.) DC.), an alternative crop for animal feed. Agron (Basel). 2024. https://doi.org/10.3390/agronomy14071433.
13. Murthazar Naim R. Growth, physiological and biochemical responses of winged bean (Psophocarpus tetragonolobus) towards different shade levels. /Murthazar Naim Raai; 2020.
14. Choi I-S, Schwarz EN, Ruhlman TA, Khiyami MA, Sabir JSM, Hajarah NH, et al. Fluctuations in Fabaceae mitochondrial genome size and content are both ancient and recent. BMC Plant Biol. 2019;19:448.
15. Wynn EL, Christensen AC. Repeats of unusual size in plant mitochondrial genomes: Identification, incidence and evolution. G3 (Bethesda). 2019;9:549–59.
16. Kubo T, Mikami T. Organization and variation of angiosperm mitochondrial genome. Physiol Plant. 2007;129:6–13.
17. Tribhuvan KU, Singh NK, Singh BK, Pandey A, Kumar S, Bishi SK et al. Chromosome-Scale, Telomere-to-Telomere Assembly of Winged Bean (Psophocarpus tetragonolobus (L.) DC.) Genome Reveals Lipid Biosynthetic Hotspots and Gene Family Dynamics. Genomics. 2025.

18. Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet. 2011;12:32–42.
19. Wright F. The effective number of codons used in a gene. Gene. 1990;87:23–9.
20. Skippington E, Barkman TJ, Rice DW, Palmer JD. Miniaturized mitogenome of the parasitic plant Viscum scurruloideum is extremely divergent and dynamic and has lost all nad genes. Proc Natl Acad Sci U S A. 2015;112:E3515–24.
21. Sloan DB, Alverson AJ, Chuckalovcak JP, Wu M, McCauley DE, Palmer JD, et al. Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PLoS Biol. 2012;10:e1001241.
22. Kubo T, Newton KJ. Plant mitochondrial genomes and their recombination behaviors. Annu Rev Plant Biol. 2008;59:561–75.
23. Adams KL, Qiu Y-L, Stoutemyer M, Palmer JD. Punctuated evolution of mitochondrial gene content: high and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. Proc Natl Acad Sci U S A. 2002;99:9905–12.
24. Wee C-C, Nor Muhammad NA, Subbiah VK, Arita M, Nakamura Y, Goh H-H. Mitochondrial genome of Garcinia mangostana L. variety Mesta. Sci Rep. 2022;12:9480.
25. Chevigny N, Schatz-Daas D, Lotfi F, Gualberto JM. DNA repair and the stability of the plant mitochondrial genome. Int J Mol Sci. 2020;21:328.
26. Liu H, Yu J, Yu X, Zhang D, Chang H, Li W, et al. Structural variation of mitochondrial genomes sheds light on evolutionary history of soybeans. Plant J. 2021;108:1456–72.
27. Sweetman C, Miller TK, Booth NJ, Shavrukov Y, Jenkins CLD, Soole KL, et al. Identification of alternative mitochondrial electron transport pathway components in chickpea indicates a differential response to salinity stress between cultivars. Int J Mol Sci. 2020;21:3844.
28. Tuteja R, Saxena RK, Davila J, Shah T, Chen W, Xiao Y-L, et al. Cytoplasmic male sterility-associated chimeric open reading frames identified by mitochondrial genome sequencing of four Cajanus genotypes. DNA Res. 2013;20:485–95.
29. He T, Ding X, Zhang H, Li Y, Chen L, Wang T, et al. Comparative analysis of mitochondrial genomes of soybean cytoplasmic male-sterile lines and their maintainer lines. Funct Integr Genomics. 2021;21:43–57.
30. Kiirika LM, Behrens C, Braun H-P, Colditz F. The mitochondrial complexome of Medicago truncatula. Front Plant Sci. 2013;4:84.
31. Maréchal A, Brisson N. Recombination and the maintenance of plant organellar genome stability. New Phytol. 2010;186:299–317.
32. Kühn K, Gualberto JM. Recombination in the stability, repair and evolution of the mitochondrial genome. In: Maréchal-Drouard L, editor. Advances in Botanical Research. Elsevier; 2012. pp. 215–52.

33. Kovar L, Nageswara-Rao M, Ortega-Rodriguez S, Mackenzie S. Gene expression and recombination activity shaped by a distinct mitochondrial genome architecture in Medicago truncatula. Plant Cell. 2018;30:1888–904.
34. Felsted RL, Egorin MJ, Leavitt RD, Bachur NR. Recombinations of subunits of Phaseolus vulgaris isolectins. J Biol Chem. 1977;252:2967–71.
35. Chen NWG, Thareau V, Ribeiro T, Magdelenat G, Ashfield T, Innes RW, et al. Common bean subtelomeres are hot spots of recombination and favor resistance gene evolution. Front Plant Sci. 2018;9:1185.
36. Bidyananda N, Jamir I, Nowakowska K, Varte V, Vendrame WA, Devi RS, et al. Plant genetic diversity studies: Insights from DNA marker analyses. Int J Plant Biol. 2024;15:607–40.
37. Merritt BJ, Culley TM, Avanesyan A, Stokes R, Brzyski J. An empirical review: Characteristics of plant microsatellite markers that confer higher levels of genetic variation. Appl Plant Sci. 2015;3:1500025.
38. Chang S, Wang Y, Lu J, Gai J, Li J, Chu P, et al. The mitochondrial genome of soybean reveals complex genome structures and gene evolution at intercellular and phylogenetic levels. PLoS ONE. 2013;8:e56502.
39. Manchekar M, Scissum-Gunn K, Song D, Khazi F, McLean SL, Nielsen BL. DNA recombination activity in soybean mitochondria. J Mol Biol. 2006;356:288–99.
40. Alverson AJ, Zhuo S, Rice DW, Sloan DB, Palmer JD. The mitochondrial genome of the legume Vigna radiata and the analysis of recombination across short mitochondrial repeats. PLoS ONE. 2011;6:e16404.
41. Shah N, Wakabayashi T, Kawamura Y, Skovbjerg CK, Wang M-Z, Mustamin Y, et al. Extreme genetic signatures of local adaptation during Lotus japonicus colonization of Japan. Nat Commun. 2020;11:253.
42. Jia W, Higgs PG. Codon usage in mitochondrial genomes: distinguishing context-dependent mutation from translational selection. Mol Biol Evol. 2008;25:339–51.
43. Iriarte A, Lamolle G, Musto H. Codon Usage Bias: An Endless Tale. J Mol Evol. 2021;89:589–93.
44. Parvathy ST, Udayasuriyan V, Bhadana V. Codon usage bias. Mol Biol Rep. 2022;49:539–65.
45. Sinha K, Jana S, Pramanik P, Bera B. Selection on synonymous codon usage in soybean (Glycine max) WRKY genes. Sci Rep. 2024;14:26530.
46. Morton BR, Wright SI. Selective constraints on codon usage of nuclear genes from Arabidopsis thaliana. Mol Biol Evol. 2007;24:122–9.
47. Singh NK, Singh BK, Giddhi A, Srivast H, Pandey A, Kumar S et al. Chloroplast genome sequencing in winged bean (Psophocarpus tetragonolobus L.) and comparative analysis with other legumes. Res Square. 2024.
48. Smith DR, Keeling PJ. Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. Proc Natl Acad Sci U S A. 2015;112:10177–84.

49. Sloan DB, Havird JC, Sharbrough J. The on-again, off-again relationship between mitochondrial genomes and species boundaries. Mol Ecol. 2017;26:2212–36.
50. Ho WK, Tanzi AS, Sang F, Tsoutsoura N, Shah N, Moore C, et al. A genomic toolkit for winged bean Psophocarpus tetragonolobus. Nat Commun. 2024;15:1901.
51. Zhao T, Zwaenepoel A, Xue J-Y, Kao S-M, Li Z, Schranz ME, et al. Whole-genome microsynteny-based phylogeny of angiosperms. Nat Commun. 2021;12:3498.
52. Gao B, Wang L, Oliver M, Chen M, Zhang J. Phylogenomic synteny network analyses reveal ancestral transpositions of auxin response factor genes in plants. Plant Methods. 2020;16:70.
53. Krishnamurthy P, Pothiraj R, Suthanthiram B, Somasundaram SM, Subbaraya U. Phylogenomic classification and synteny network analyses deciphered the evolutionary landscape of aldo-keto reductase (AKR) gene superfamily in the plant kingdom. Gene. 2022;816:146169.
54. Niu Y, Gao C, Liu J. Complete mitochondrial genomes of three Mangifera species, their genomic structure and gene transfer from chloroplast genomes. BMC Genomics. 2022;23:147.
55. Lu G, Zhang K, Que Y, Li Y. Assembly and analysis of the first complete mitochondrial genome of Punica granatum and the gene transfer from chloroplast genome. Front Plant Sci. 2023;14:1132551.
56. Park H-S, Lee WK, Lee S-C, Lee HO, Joh HJ, Park JY, et al. Inheritance of chloroplast and mitochondrial genomes in cucumber revealed by four reciprocal F1 hybrid combinations. Sci Rep. 2021;11:2506.
57. Choi I-S, Wojciechowski MF, Ruhlman TA, Jansen RK. In and out: Evolution of viral sequences in the mitochondrial genomes of legumes (Fabaceae). Mol Phylogenet Evol. 2021;163:107236.
58. Lee C, Choi I-S, Cardoso D, de Lima HC, de Queiroz LP, Wojciechowski MF, et al. The chicken or the egg? Plastome evolution and an independent loss of the inverted repeat in papilionoid legumes. Plant J. 2021;107:861–75.
59. Choi I-S, Cardoso D, de Queiroz LP, de Lima HC, Lee C, Ruhlman TA, et al. Highly Resolved Papilionoid Legume Phylogeny Based on Plastid Phylogenomics. Front Plant Sci. 2022;13:823190.
60. Satrio RD, Nikmah IA, Fendiyanto MH, Pratami MP, Dewi AP, Hariri MR et al. A complete chloroplast and mitochondrial genome for velvet bean (Mucuna pruriens, Fabaceae), with genome structure and intergenomic sequence transfers analyses. 2023.
61. Yang S, Li G, Li H. Molecular characterizations of genes in chloroplast genomes of the genus Arachis L. (Fabaceae) based on the codon usage divergence. PLoS ONE. 2023;18:e0281843.
62. Diamond J. Evolution, consequences and future of plant and animal domestication. Nature. 2002;418:700–7.
63. Alam O, Purugganan MD. Domestication and the evolution of crops: variable syndromes, complex genetic architectures, and ecological entanglements. Plant Cell. 2024;36:1227–41.
64. Birch JL, Cantrill DJ, Walsh NG, Murphy DJ. Phylogenetic investigation and divergence dating of*Poa*(Poaceae, tribe Poeae) in the Australasian region: Australasian*Poa*Phylogenetics and Dating. Bot J Linn Soc. 2014;175:523–52.

65. Ilic K, SanMiguel PJ, Bennetzen JL. A complex history of rearrangement in an orthologous region of the maize, sorghum, and rice genomes. Proc Natl Acad Sci U S A. 2003;100:12265–70.
66. Chen Q, Li W, Tan L, Tian F. Harnessing knowledge from maize and rice domestication for new crop breeding. Mol Plant. 2021;14:9–26.
67. Wang Y, Jiang D, Guo K, Zhao L, Meng F, Xiao J, et al. Comparative analysis of codon usage patterns in chloroplast genomes of ten Epimedium species. BMC Genom Data. 2023;24:3.
68. Sun M, Wang J, Smagghe G, Dai R, Wang X, Yang Y et al. Description of mitochondrial genomes and phylogenetic analysis of Megophthalminae (Hemiptera: Cicadellidae). J Insect Sci. 2024;24.
69. Indrabalan UB, Suresh KP, Shivamallu C, Patil SS. An extensive evaluation of codon usage pattern and bias of structural proteins p30, p54 and, p72 of the African swine fever virus (ASFV). Virusdisease. 2021;32:810–22.
70. Xiao M, Hu X, Li Y, Liu Q, Shen S, Jiang T, et al. Comparative analysis of codon usage patterns in the chloroplast genomes of nine forage legumes. Physiol Mol Biol Plants. 2024;30:153–66.
71. Rao A, Chen Z, Wu D, Wang Y, Hou N, Afr. J Bot. 2024;164:429–37.
72. Li Y, Hu X, Xiao M, Huang J, Lou Y, Hu F, et al. An analysis of codon utilization patterns in the chloroplast genomes of three species of Coffea. BMC Genom Data. 2023;24:42.
73. Wang L, Xing H, Yuan Y, Wang X, Saeed M, Tao J, et al. Genome-wide analysis of codon usage bias in four sequenced cotton species. PLoS ONE. 2018;13:e0194372.
74. Paul P, Malakar AK, Chakraborty S. Codon usage vis-a-vis start and stop codon context analysis of three dicot species. J Genet. 2018;97:97–107.
75. Qiu S, Zeng K, Slotte T, Wright S, Charlesworth D. Reduced efficacy of natural selection on codon usage bias in selfing Arabidopsis and Capsella species. Genome Biol Evol. 2011;3:868–80.
76. Kimura M. Possibility of extensive neutral evolution under stabilizing selection with special reference to nonrandom usage of synonymous codons. Proc Natl Acad Sci U S A. 1981;78:5773–7.
77. Hallee L, Khomtchouk BB. Machine learning classifiers predict key genomic and evolutionary traits across the kingdoms of life. Sci Rep. 2023;13:1–14.
78. Zavala B, Dineen L, Fisher KJ, Opulente DA, Harrison M-C, Wolters JF et al. Genomic factors shaping codon usage across the Saccharomycotina subphylum. G3 Genes|Genomes|Genetics. 2024;14:jkae207.
79. Yannai A, Katz S, Hershberg R. The Codon Usage of Lowly Expressed Genes Is Subject to Natural Selection. Genome Biol Evol. 2018;10:1237–46.
80. Wang Z, Xu B, Li B, Zhou Q, Wang G, Jiang X, et al. Comparative analysis of codon usage patterns in chloroplast genomes of six species. PeerJ. 2020;8:e8251.
81. Allen JF. Why chloroplasts and mitochondria retain their own genomes and genetic systems: Colocation for redox regulation of gene expression. Proc Natl Acad Sci U S A. 2015;112:10231–8.
82. Wang D, Wu Y-W, Shih AC-C, Wu C-S, Wang Y-N, Chaw S-M. Transfer of chloroplast genomic DNA to mitochondrial genome occurred at least 300 MYA. Mol Biol Evol. 2007;24:2040–8.

83. Cui H, Ding Z, Zhu Q, Wu Y, Qiu B, Gao P. Comparative analysis of nuclear, chloroplast, and mitochondrial genomes of watermelon and melon provides evidence of gene transfer. Sci Rep. 2021;11:1595.
84. Zhang T, Fang Y, Wang X, Deng X, Zhang X, Hu S, et al. The complete chloroplast and mitochondrial genome sequences of Boea hygrometrica: insights into the evolution of plant organellar genomes. PLoS ONE. 2012;7:e30531.
85. Li G, Zhang H, Lin Z, Li H, Xu G, Xu Y, et al. Comparative analysis of chloroplast and mitochondrial genomes of sweet potato provides evidence of gene transfer. Sci Rep. 2024;14:4547.
86. Doyle J. DNA Protocols for Plants. Molecular Techniques in Taxonomy. Berlin, Heidelberg: Springer Berlin Heidelberg; 1991. pp. 283–93.
87. Andrews S. FastQC: A quality control analysis tool for high throughput sequencing data. Github.
88. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
89. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019;37:1155–62.
90. Uliano-Silva M, Ferreira JGRN, Krasheninnikova K, Darwin Tree of Life Consortium, Formenti G, Abueg L, et al. MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads. BMC Bioinformatics. 2023;24:288.
91. Singh NK, Singh BK, Giddhi A, Srivastava H, Pandey A, Kumar S et al. Chloroplast genome sequencing in winged bean (Psophocarpus tetragonolobus L.) and comparative analysis with other legumes. 2024.
92. Jin J-J, Yu W-B, Yang J-B, Song Y, dePamphilis CW, Yi T-S, et al. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol. 2020;21:241.
93. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome assemblies. Bioinformatics. 2015;31:3350–2.
94. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, et al. GeSeq - versatile and accurate annotation of organelle genomes. Nucleic Acids Res. 2017;45:W6–11.
95. Greiner S, Lehwark P, Bock R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Res. 2019;47:W59–64.
96. Du L, Zhang C, Liu Q, Zhang X, Yue B, Hancock J. Krait: an ultrafast tool for genome-wide survey of microsatellites and primer design. Bioinformatics. 2018;34:681–3.
97. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27:573–80.
98. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 2001;29:4633–42.

99. Katoh K, Misawa K, Kuma K-I, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30:3059–66.
100. Nakamura T, Yamada KD, Tomii K, Katoh K. Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics. 2018;34:2490–2.
101. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 2019;35:4453–5.
102. Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics. 2014;30:i541–8.
103. Zhang C, Rabiee M, Sayyari E, Mirarab S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics. 2018;19:153.
104. Guy L, Kultima JR, Andersson SGE. genoPlotR: comparative gene and genome visualization in R. Bioinformatics. 2010;26:2334–5.
105. Zhang Z. KaKs_Calculator 3.0: Calculating selective pressure on coding and non-coding sequences. Genomics Proteom Bioinf. 2022;20:536–40.
106. Álvarez-Carretero S, Kapli P, Yang Z. Beginner's Guide on the Use of PAML to Detect Positive Selection. Mol Biol Evol. 2023;40.
107. Jeffares DC, Tomiczek B, Sojo V, dos Reis M. A Beginners Guide to Estimating the Non-synonymous to Synonymous Rate Ratio of all Protein-Coding Genes in a Genome. Parasite Genomics Protocols. 2015;:65–90.
108. Kuhn M. Building Predictive Models in R Using the caret Package. J Stat Soft. 2008;28:1–26.
109. Panda B, Bisoyi SS, Panigrahy S, Mohanty P. Machine learning techniques for imbalanced multiclass malware classification through adaptive feature selection. PeerJ Comput Sci. 2025;11:e2752.
110. Brownlee J. XGBoost With Python: Gradient Boosted Trees with XGBoost and scikit-learn. Machine Learning Mastery; 2016.
111. Wang S, Zhao S, Lv L, Xie X, Cui T, Yao Q, et al. Color imaging system high dynamic range colorimetric characterization modeling based on fusion kernel XGBoost. J Opt Soc Am Opt Image Sci Vis. 2025;42:908–17.
112. Wickham H. Ggplot2: elegant graphics for data analysis. New York: Springer-; 2016.
113. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
114. Darzentas N. Circoletto: visualizing sequence similarity with Circos. Bioinformatics. 2010;26:2620–1.

Tables

Tables 2 and 3 are available in the Supplementary Files section.

Figures

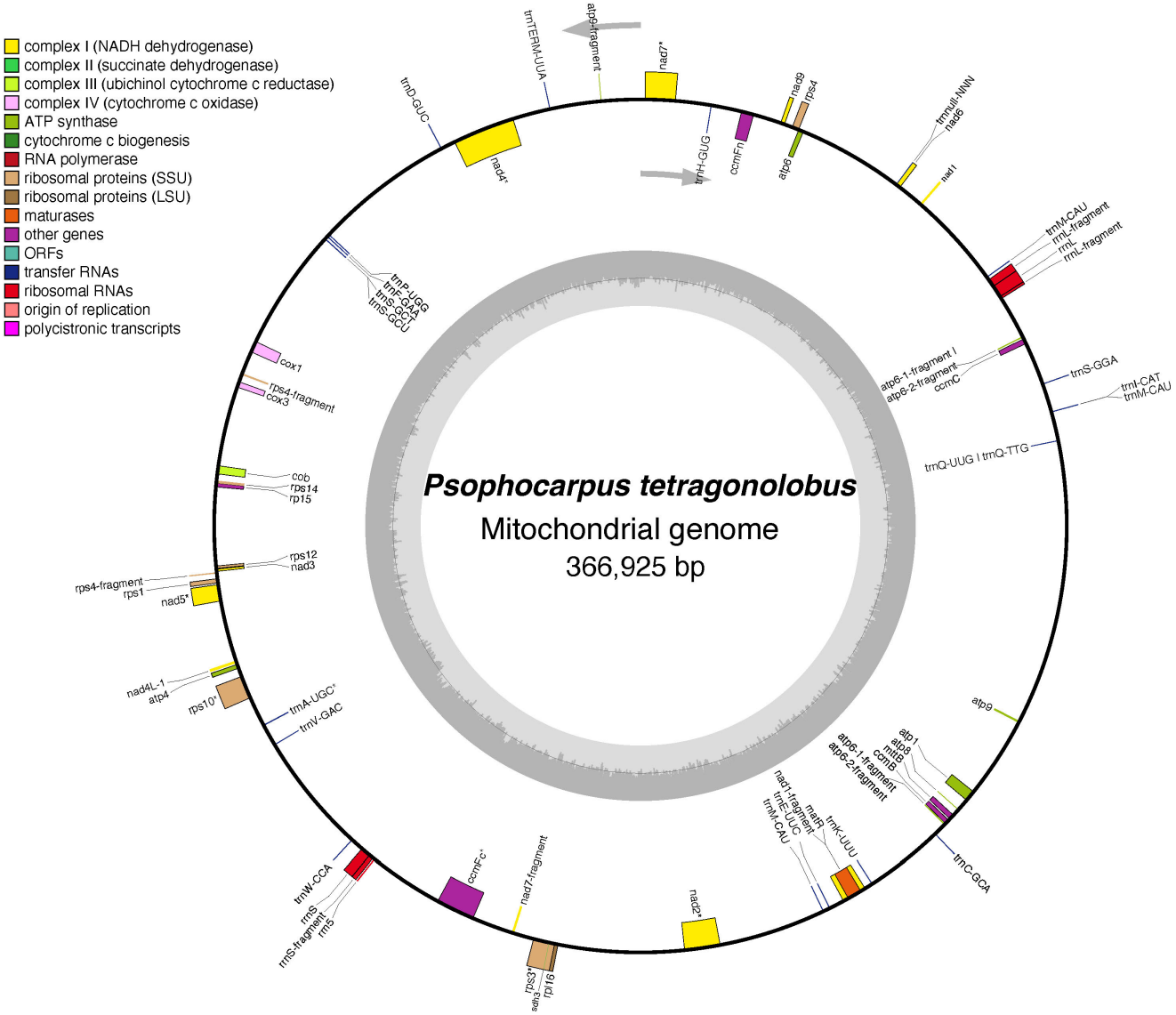

Figure 1

Annotated mitochondrial genome of Psophocarpus tetragonolobus.The circular diagram represents the complete mitochondrial genome spanning 366,925 bp. Key functional gene categories are color-coded, including polycistronic transcripts, ribosomal RNAs, transfer RNAs, maturases, ribosomal proteins (LSU and SSU), RNA polymerase subunits, and components of respiratory chain complexes I-IV, ATP synthase, and cytochrome c biogenesis. Additional details, such as partial or fragmented genes, are labeled where

applicable. The outermost circle displays the genomic arrangement with specified gene loci and their orientations.

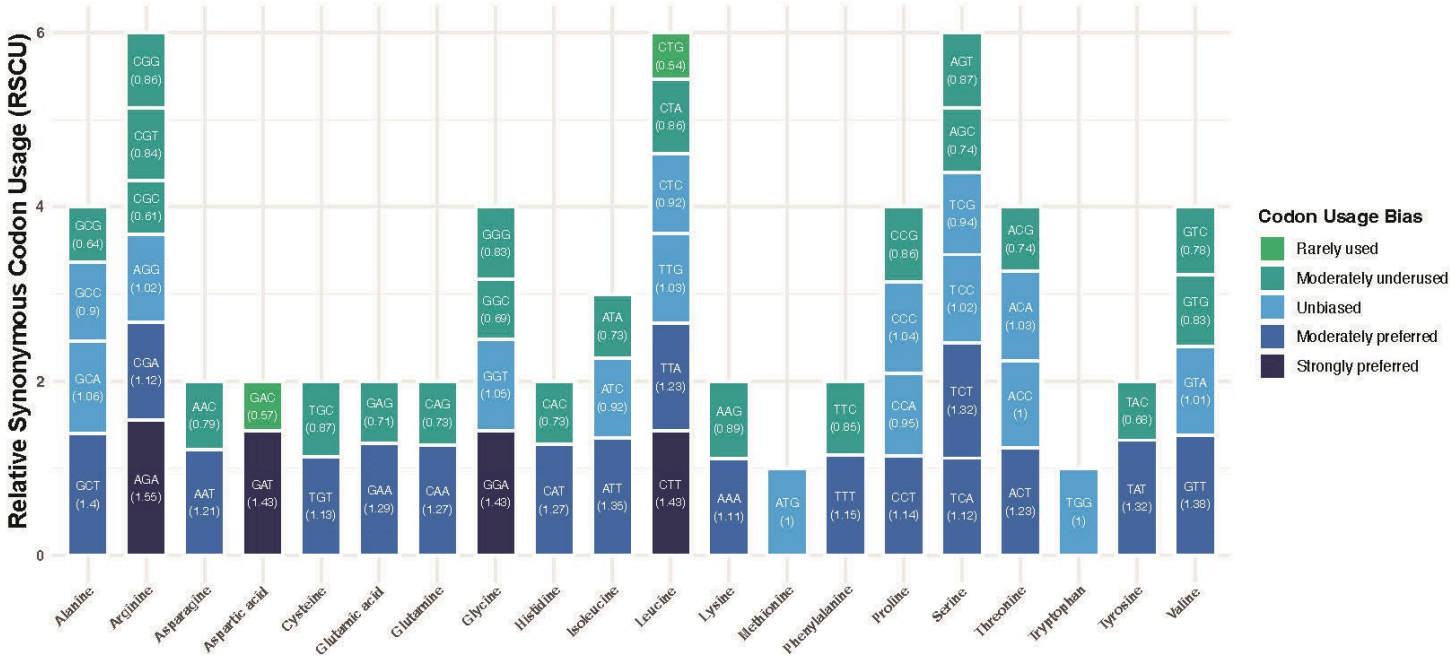

Figure 2

Codon usage and relative synonymous codon usage (RSCU) in the *Psophocarpus tetragonolobus* mitochondrial genome. This figure presents a stacked barplot showing Relative Synonymous Codon Usage (RSCU) values for each amino acid. Each bar represents one amino acid, and is subdivided into its synonymous codons. Segment heights correspond to codon-specific RSCU values, with each segment labeled by codon and its RSCU (rounded to two decimal places). Codons are color-coded based on usage bias categories: strongly preferred (RSCU > 1.4, darkest blue), moderately preferred (1.1–1.4), unbiased (0.9–1.1), moderately underused (0.6–0.9), and rarely used (<0.6, lightest blue). This visualization highlights codon preference patterns across amino acid families and reflects the codon bias present in mitochondrial genes, potentially shaped by evolutionary and translational constraints

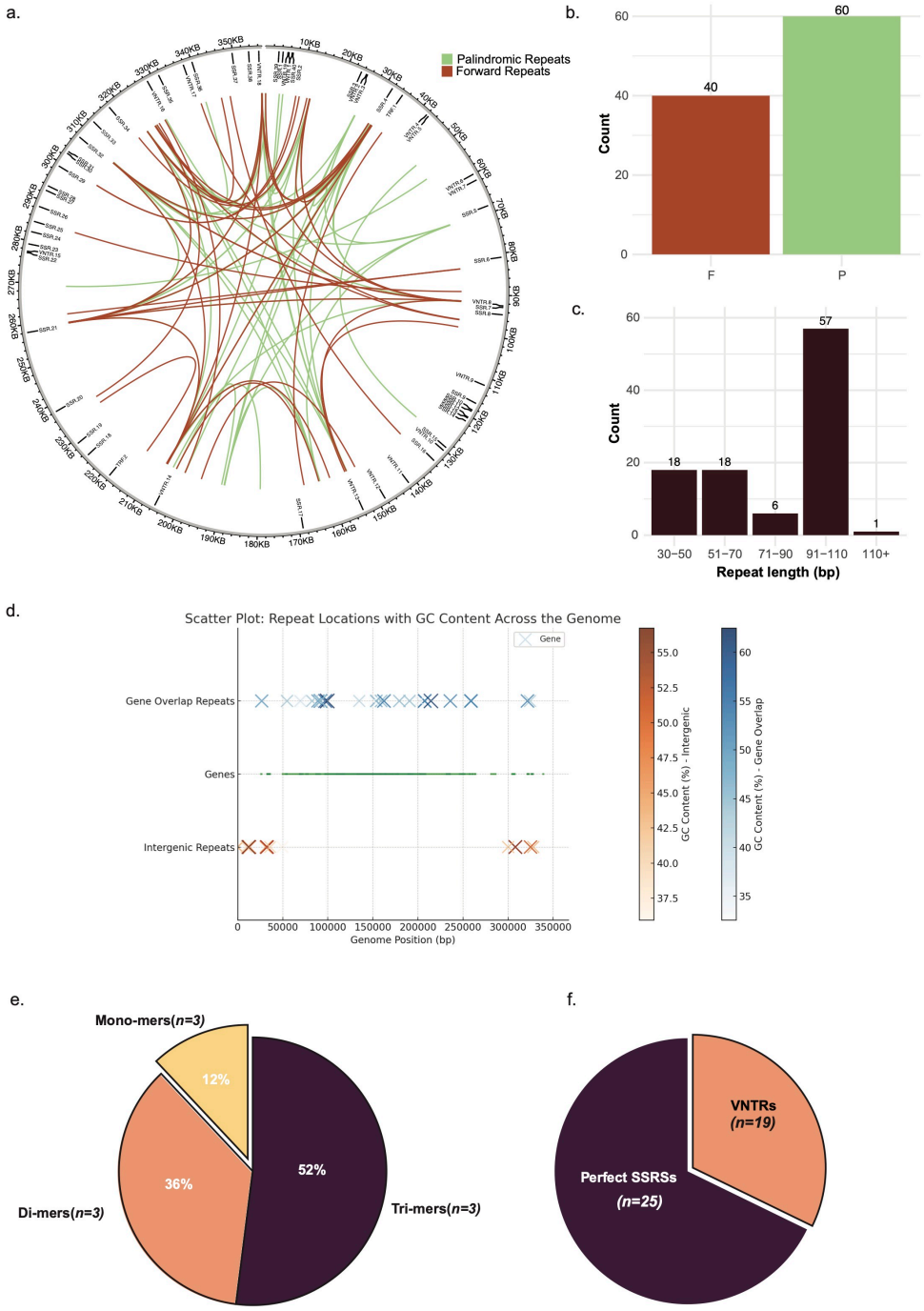

Figure 3

Repeat analysis of the Psophocarpus tetragonolobus mitochondrial genome.(a) Repeat summary illustrating the distribution of repeat types across the genome, including forward, palindromic, and tandem repeats, categorized by length (bp). (b) Distribution of short sequence repeats (SSRs) and variable number tandem repeats (VNTRs) across genome positions. (c) Percentage breakdown of repeat types in the genome, showing the proportions of perfect SSRs, VNTRs, and other repeat categories. (d-f)

Detailed characterization of repeat units by size, including mono-, di-, and tri-mers. Repeat loci and their frequencies are mapped to specific regions of the genome.

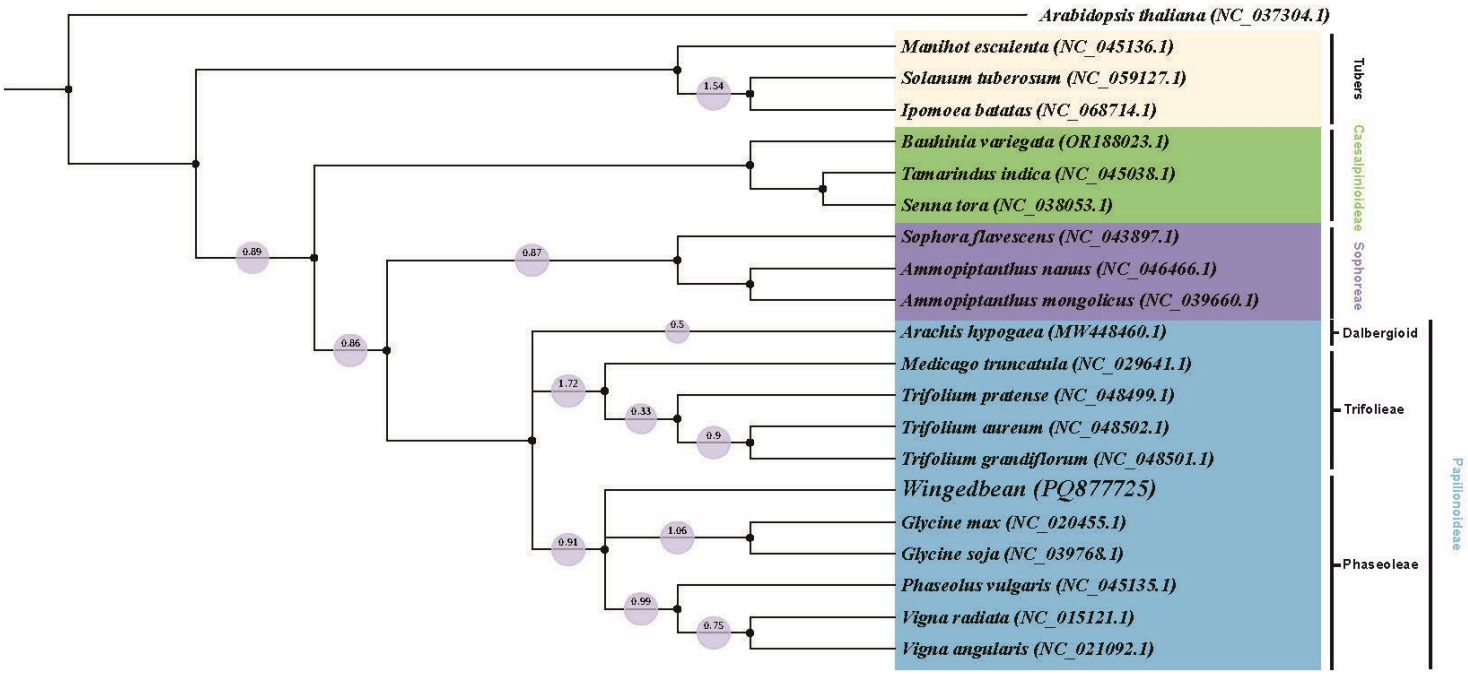

Figure 4

Phylogenetic analysis of *Psophocarpus tetragonolobus* based on mitochondrial genome sequences. Phylogenetic analysis of Psophocarpus tetragonolobus based on mitochondrial genome sequences. The tree illustrates the evolutionary relationships of P. tetragonolobus with other legume species, including Vigna angularis, Vigna radiata, Phaseolus vulgaris, Glycine max, Glycine soja, Trifolium grandiflorum, Trifolium aureum, Trifolium pratense, Medicago truncatula, and Arachis hypogaea. Additional legume representatives from Sophoreae (Sophora flavescens), Caesalpinioideae (Senna tora, Tamarindus indica, Bauhinia variegata), and Dalbergioid groups are included. Tuber crops (Ipomoea batatas, Solanum tuberosum, and Manihot esculenta) are analyzed for comparison. Arabidopsis thaliana serves as the outgroup to root the tree. Bootstrap values are indicated at major nodes. The analysis confirms the placement of P. tetragonolobus within the Phaseoleae tribe of Papilionoideae. The scale bar represents genetic distance.

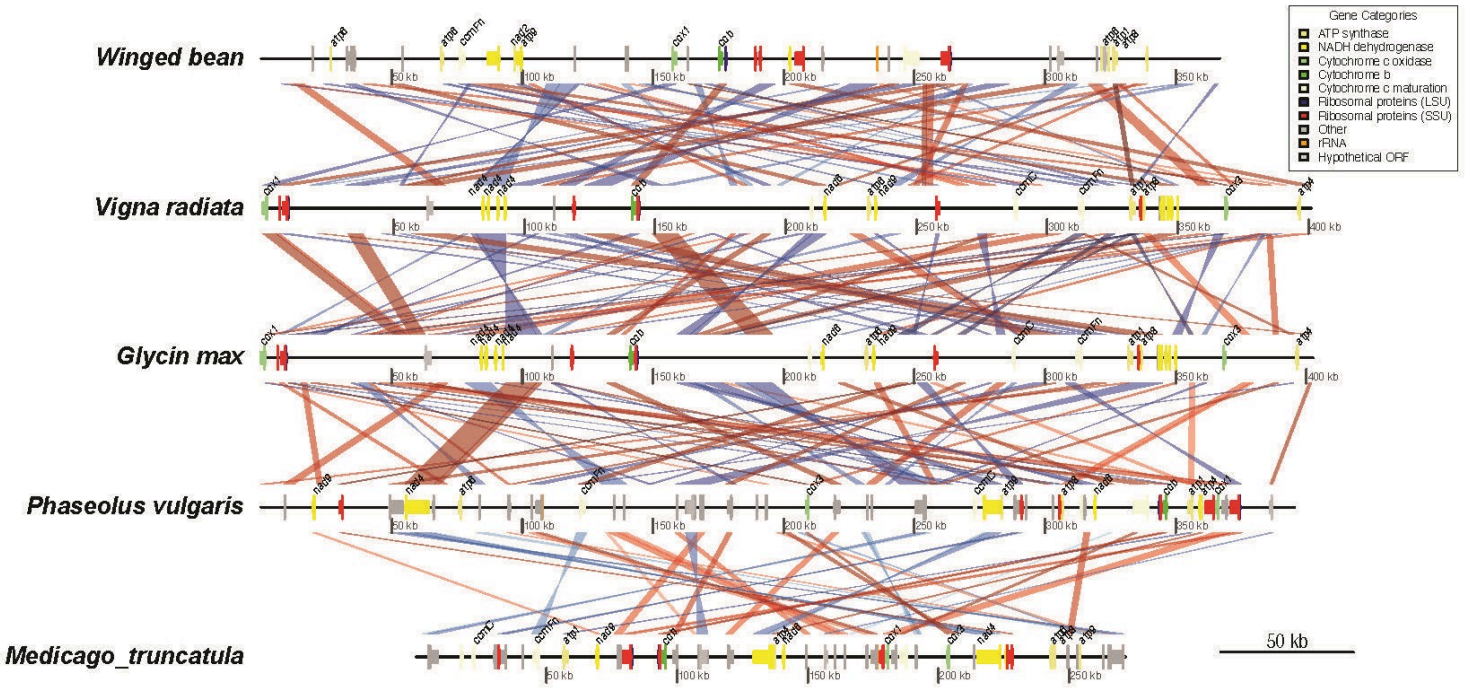

Figure 5

Comparative synteny analysis of the Psophocarpus tetragonolobus mitochondrial genome. Comparative synteny analysis of the mitochondrial genome of Psophocarpus tetragonolobus with related legumes. The linear tracks represent the mitochondrial genomes of *P. tetragonolobus (winged bean), G. max, V. radiata, P. vulgaris, and M. truncatula.* Conserved syntenic blocks between species are indicated by colored connectors, illustrating homologous genomic regions. Gene directionality is represented by arrows (strand orientation details provided in Supplementary Table S3). Core mitochondrial genes involved in respiration, cytochrome maturation, and ribosomal functions are highlighted. Presence and absence patterns of mitochondrial genes across species are summarized in Supplementary Figure S6. Major inversions, rearrangements, and lineage-specific deletions are visualized through interruptions or changes in block orientation, revealing both conserved and dynamic structural features within Fabaceae mitochondrial genomes.

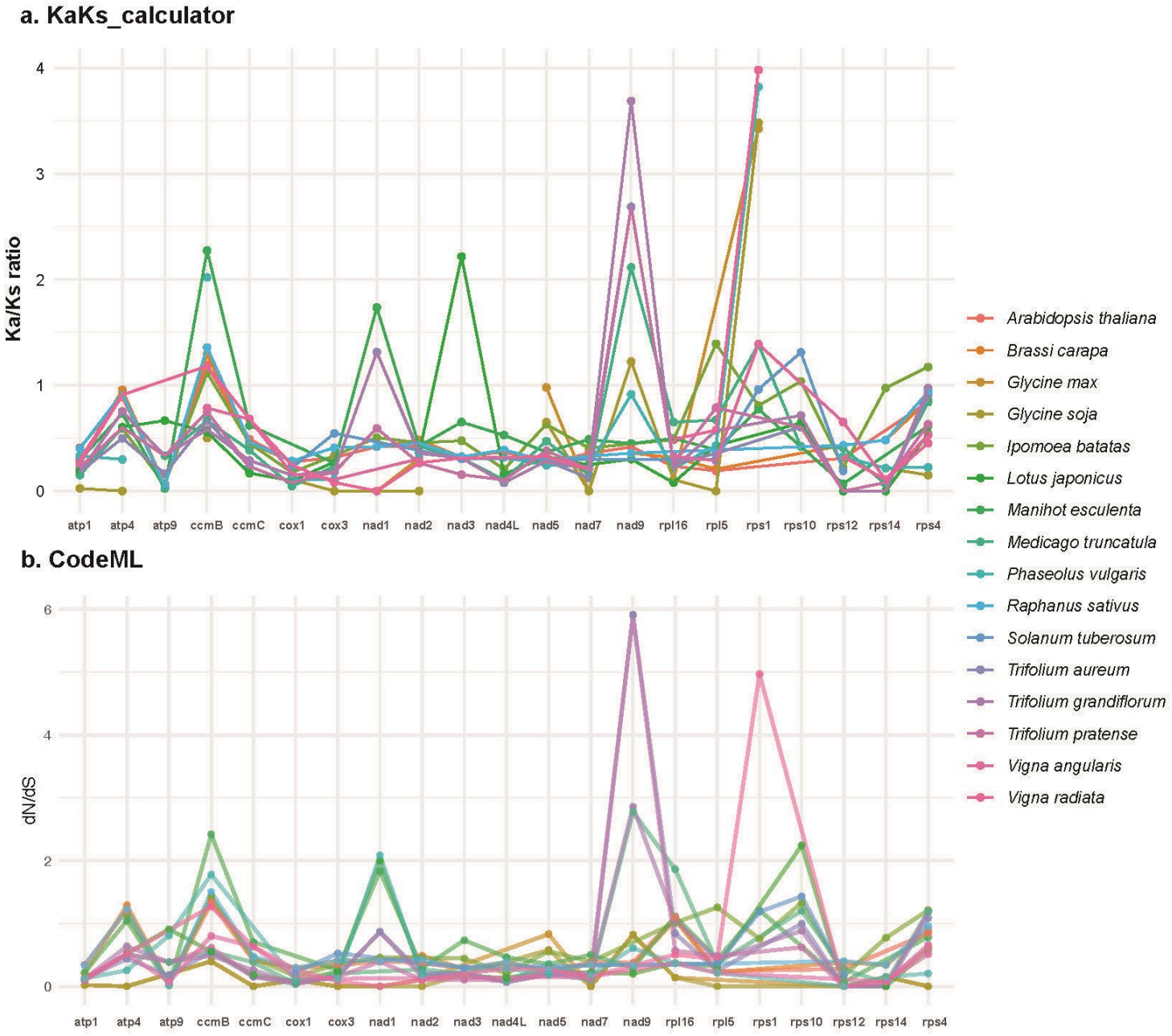

Figure 6

Ka/Ks (dN/dS) ratio analysis of protein-coding genes in the mitochondrial genome of *P. tetragonolobus*. (a) Ka/Ks ratios estimated using KaKs_Calculator3, providing pairwise substitution rate comparisons between P. tetragonolobus and related plant species (Arabidopsis thaliana, Glycine max, Medicago truncatula, Vigna radiata, and others) (b) Phylogenetically informed dN/dS ratios estimated using codeml (PAML package), accounting for evolutionary relationships across the same set of genes and species. Each dot represents a specific gene's Ka/Ks ratio, reflecting selective pressures: values <1 indicate purifying selection, values ≈1 suggest neutral evolution, and values >1 imply positive selection. Comparative analysis highlights variation in evolutionary constraints across mitochondrial genes among legumes and other angiosperms.

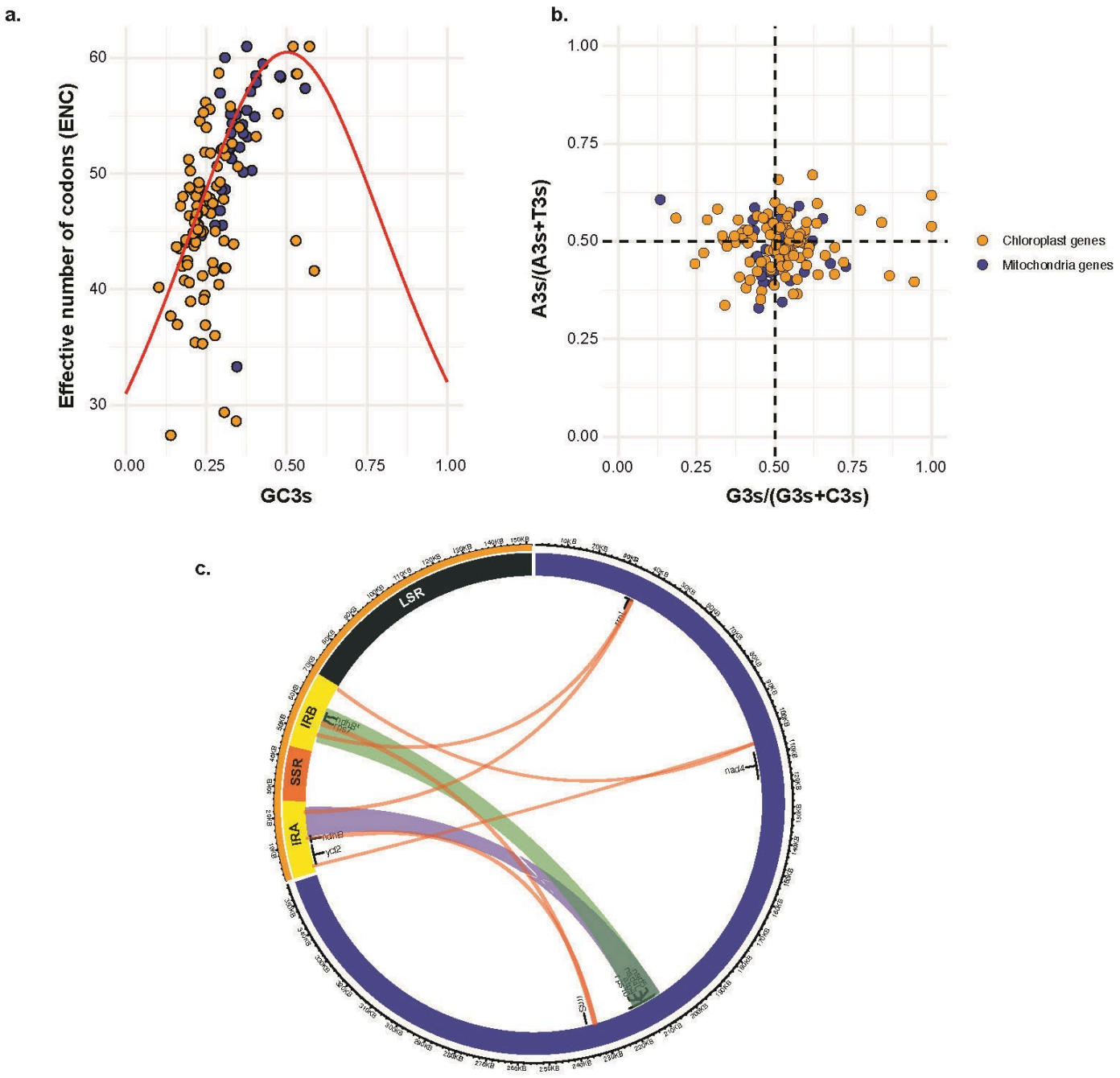

Figure 7

Codon usage bias and genomic features of Psophocarpus tetragonolobus mitochondrial and chloroplast genes.(a) The relationship between GC content at the third codon position (GC3s) and the effective number of codons (ENC) shows a codon usage bias for mitochondrial and chloroplast genes. (b) GC content ratios (G3s/(G3s+C3s) and A3s/(A3s+T3s)) for mitochondrial and chloroplast genes, highlighting compositional biases. (c) Genomic distribution of key genes, including mitochondrial genes

(nad4, nad5, rps10) and chloroplast genes (ycf2, ndhB, rps7), along with structural features like inverted repeats (IRA, IRB).

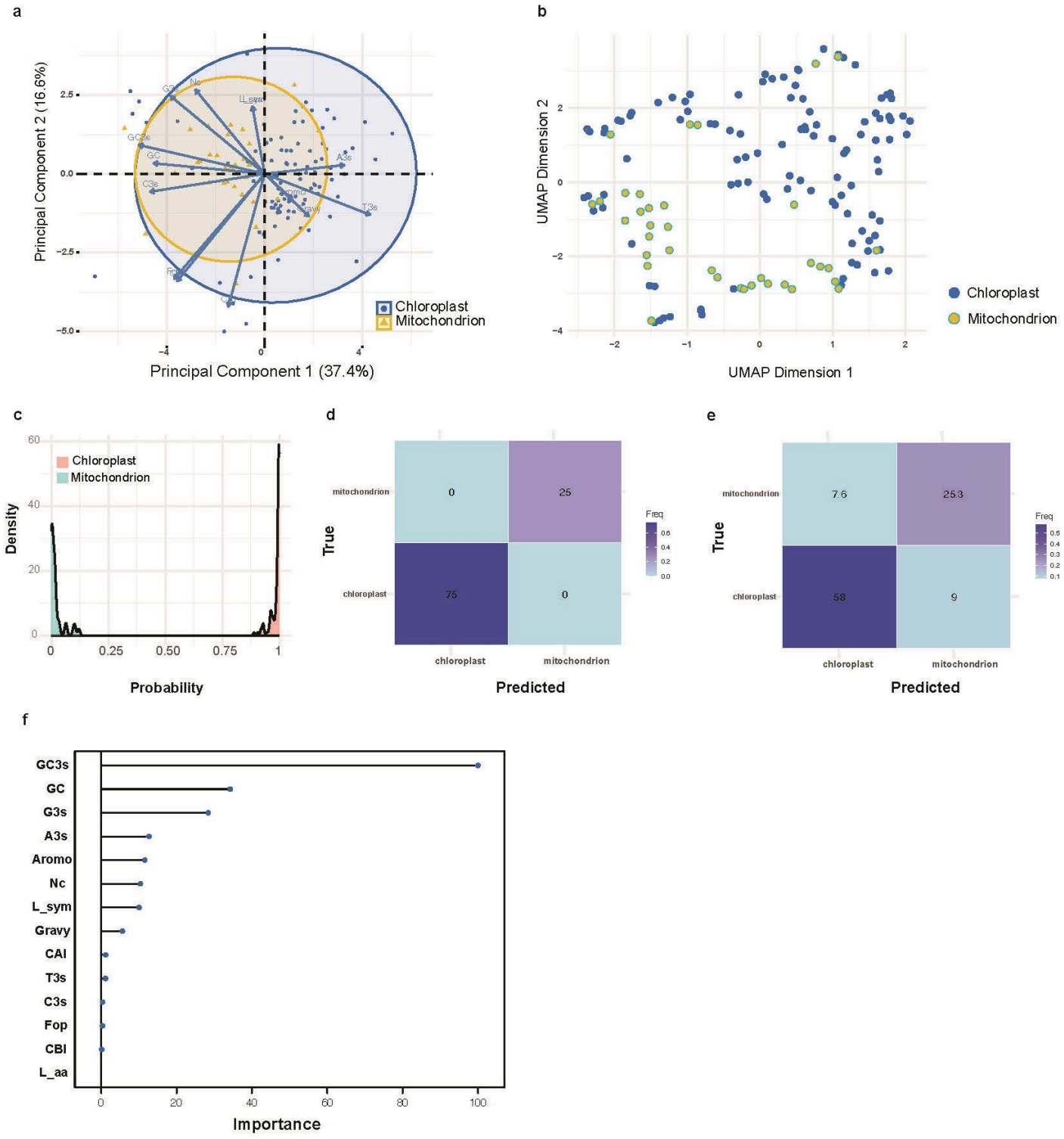

Figure 8

Machine learning–based classification of mitochondrial and chloroplast genes using codon usage bias metrics in *P. tetragonolobus*. (a) Principal Component Analysis (PCA) plot showing separation of mitochondrial (orange) and chloroplast (blue) genes based on 14 codon usage–related features, with

PC1 and PC2 explaining 37.4% and 16.6% of the variance, respectively. (b) Uniform Manifold Approximation and Projection (UMAP) embedding of the same features reveals distinct non-linear clustering of organelle gene origins. (c) Confusion matrix for the optimal classification model (xgbTree), demonstrating high accuracy in predicting gene origin, with most chloroplast and mitochondrial genes correctly classified. (d) Bar plot of class distribution for true vs. predicted labels, indicating strong model performance across both classes.

(e) Probability density curves of predicted probabilities for each class, illustrating clear separation between mitochondrial and chloroplast genes. (f) Variable importance ranking from the xgbTree model, highlighting GC content at the third codon position (GC3s) as the most influential feature, followed by overall GC content (GC), G3s, and A3s. Additional features include CAI, CBI, Fop, Nc, L_sym, L_aa, GRAVY, aromaticity (Aromo), and nucleotide frequencies (A3s, T3s, C3s, G3s), collectively enabling robust classification (AUC = 0.96).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table2.xlsx
- Table3.xlsx
- SupplementaryFigure.docx
- SuppTableS1.xlsx
- SuppTableS2.xlsx
- SuppTableS3.xls
- SuppTableS4.xlsx