Scientific Reports

https://doi.org/10.1038/s41598-026-66878-3

Article in Press

Decoding the chloroplast genome of the Amazonian fruit *Myrciaria Dubia* (camu-camu): molecular evolution and comparative genomics across the order Myrtales

Received: 25 August 2025

Accepted: 11 August 2026

Published online: 14 August 2026

Cite this article as: Castro J.C., Leonardo D.A., Cobos M. *et al.* Decoding the chloroplast genome of the Amazonian fruit *Myrciaria Dubia* (camu-camu): molecular evolution and comparative genomics across the order Myrtales. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-66878-3

Juan C. Castro, Diego A. Leonardo, Marianela Cobos, Jhon A. Vargas, Jae D. Paredes, Pedro M. Adrianzen, Carlos G. Castro, Sixto A. Imán, Hicler N. Rodriguez, Jorge L. Marapara & J. Dylan Maddox

Decoding the Chloroplast Genome of the Amazonian Fruit *Myrciaria dubia* (camu-camu): Molecular Evolution and Comparative Genomics Across the Order Myrtales

Juan C. Castro[1,2*], Diego A. Leonardo[3], Marianela Cobos[1,2], Jhon A. Vargas[3], Jae D. Paredes[1], Pedro M. Adrianzen[1,2], Carlos G. Castro[1], Sixto A. Imán[4,5], Hicler N. Rodriguez[1,2], Jorge L. Marapara[1,2], J. Dylan Maddox[6*]

[1] Specialized Unit of Biotechnology Research Laboratory, Natural Resources Research Center of UNAP, Universidad Nacional de la Amazonía Peruana (UNAP), Iquitos 16001, Peru
[2] Academic Department of Biomedical Sciences and Biotechnology, Faculty of Biological Sciences, Universidad Nacional de la Amazonía Peruana (UNAP), Iquitos 16001, Peru
[3] Institute of Physics of São Carlos, Universidade de São Paulo, Street João Dagnone 1100, São Carlos, SP 13563-723, Brazil
[4] Academic Department of Ecology and Conservation, Faculty of Forestry Sciences, Universidad Nacional de la Amazonía Peruana (UNAP), Iquitos 16001, Peru
[5] Direction of Genetic Resources and Biotechnology, San Roque Agricultural Experimental Station, Instituto Nacional de Innovación Agraria (INIA), Iquitos 16001, Peru
[6] Pritzker Laboratory for Molecular Systematics and Evolution, Field Museum of Natural History, 1400 S. Lake Shore Drive, Chicago, IL 60605, USA
* Authors to whom correspondence should be addressed:
juan.castro@unapiquitos.edu.pe; Tel.: +51-965-017215
dmaddox@fieldmuseum.org; Tel.: +1 (312) 665-7134

Abstract

Myrciaria dubia (camu-camu) is an economically important Amazonian fruit species with severely limited genomic resources despite growing commercial interest and conservation needs. This study sequenced, assembled, and comprehensively analyzed the complete chloroplast (cp) genome of *M. dubia* and performed comparative genomic analyses across the order Myrtales, including codon usage, codon decoding mechanisms, repetitive elements, nucleotide diversity, and phylogenetic reconstruction. The cp genome spanned 158,664 bp with a typical quadripartite organization: a large single-copy region (LSC = 87,640 bp), a small single-copy region (SSC = 18,324 bp), and paired inverted repeats (IRA/IRB = 26,350 bp each). Functional annotation identified 135 genomic features, comprising 84 protein-coding genes, 37 tRNA genes, eight rRNA genes, and five pseudogenes. Codon usage analysis revealed a strong AT bias (62.50%), with 30 preferred codons (RSCU >1.0) predominantly ending with A/T nucleotides. Codon decoding analysis assigned all 61 sense codons to five mechanistic categories, collectively achieving complete codon coverage across all 23 Myrtales accessions. Repetitive element analysis identified 69 simple sequence repeats in *M. dubia* (91.3% mononucleotide, predominantly adenine motifs in LSC intergenic spacers) and 439 long dispersed repeat pairs (97.5% palindromic, concentrated in the inverted repeat regions). Nucleotide diversity varied significantly across structural regions (SSC, mean π = 0.173; LSC, mean π = 0.084; IRA/IRB, mean π = 0.010–0.015), with the highest diversity window (π = 0.240) located in the SSC at approximately 117.7 kb. SWAN analysis of 78 protein-coding genes identified *rpl22*, *ycf1*, *matK*, *ndhF*, and *ndhD* as the most variable loci, with no gene exceeding ω = 1.0, indicating a pervasive purifying selection. Comparative analysis across 19 Myrtaceae accessions demonstrated complete gene collinearity and conserved IR boundary arrangement, with divergence concentrated in the SSC intergenic regions. Phylogenetic analysis positioned *M. dubia* as a sister species to *Plinia trunciflora* within the Myrteae tribe, nested in a well-supported monophyletic Myrtaceae clade (bootstrap = 100%). These findings provide the first complete cp genome resource for *M. dubia* and establish an essential foundation for conservation genetics, breeding programs, and evolutionary studies of this economically valuable Amazonian species.

Keywords: Camu-camu, Chloroplast Genome, Genetic Variation, Microsatellite Repeats, *Myrciaria dubia*, Myrtaceae, Phylogeny.

Introduction

Chloroplast (cp) genomes are valuable genetic resources for understanding plant evolution and developing molecular tools for species identification and conservation (1,2). These circular DNA molecules typically range from 120 to 200 kb and contain genes essential for photosynthesis, transcription, and translation [3]. The conserved four-part structure of most cp genomes, consisting of large single-copy (LSC) and small single-copy (SSC) regions separated by two inverted repeat (IR) regions, provides stability for essential functions and sufficient variation for evolutionary studies [4,5]. Recent advances in DNA sequencing technologies have made cp genome sequencing more accessible, leading to a rapid increase in the number of available genome sequences [4,6]. These resources have

proven particularly valuable for resolving evolutionary relationships in complex plant families and developing molecular markers for conservation applications [5,7]. Comparative chloroplast genomics has become a standard approach for understanding genome evolution, identifying variable regions for phylogenetic studies, and developing species-specific molecular markers [8–11].

The family Myrtaceae is one of the largest plant families, with approximately 5,950 species distributed primarily in tropical and subtropical regions [12,13]. This family includes many economically important species valued for their fruits, essential oils, and medicinal properties [14]. Despite their economic importance, many Myrtaceae species remain poorly characterized at the genomic level, particularly those from tropical regions with high biodiversity [15,16]. Within Myrtaceae, the genus *Myrciaria* has gained increasing scientific attention due to its remarkable nutritional and therapeutic properties, particularly the species *Myrciaria dubia* (Kunth) McVaugh, commonly known as camu-camu [17,18]

M. dubia is a fruit-bearing shrub native to the Amazon rainforest. This species has gained global attention because it contains the highest documented concentrations of vitamin C among natural food sources, with levels reaching 2,000-3,000 mg per 100 g of fresh fruit [19,20]. In addition to vitamin C, *M. dubia* fruits contain significant amounts of other antioxidant compounds, including anthocyanins, flavonoids, and phenolic acids [19,21–23]. These properties have generated substantial commercial interest in its use as a functional food and nutraceutical ingredient [24,25].

Despite its economic importance and unique nutritional profile, *M. dubia* has been poorly represented in genomic databases. Previous molecular studies have focused on specific gene regions, leaving substantial gaps in our understanding of genome organization and evolutionary relationships [26,27]. The lack of complete cp genome information limits the development of molecular markers for conservation planning and breeding programs. In addition, comparative chloroplast genomics has proven effective for resolving relationships within large plant families, where traditional morphological characteristics may be insufficient [7,28]. Analysis of genome structure, gene content, and sequence variation provides insights into evolutionary processes and identifies regions suitable for phylogenetic reconstruction [5,9].

The present study addresses critical knowledge gaps by providing the first complete cp genome sequence and a comprehensive comparative analysis of *M. dubia* within the order Myrtales. Specifically, the objectives were as follows: (1) to characterize genome organization, structural architecture, and gene content; (2) to analyze synonymous codon usage bias and codon decoding mechanisms across five mechanistic categories; (3) to identify and characterize simple sequence repeats and long dispersed repeats; (4) to assess nucleotide diversity and evolutionary rate variation at both genome-wide and protein-coding gene levels; (5) to conduct comparative genomic analyses of IR boundary architecture, gene collinearity, and sequence conservation across 23 Myrtales; and (6) to reconstruct phylogenetic relationships within Myrtales using complete cp genome sequences. Together, these analyses provide essential genomic resources for future research on this economically important Amazonian species and contribute to a broader understanding of plastid genome evolution, codon biology, and molecular systematics within the Myrtaceae family.

Results

Genome organization and gene distribution patterns

The complete cp genome of *M. dubia* was successfully assembled as a circular double-stranded DNA molecule with a total length of 158,664 bp (Fig. 1). The genome exhibited a characteristic quadripartite architecture comprising four structurally and functionally distinct regions. The large single-copy (LSC) region was the predominant genomic component at 87,640 bp, accounting for 55.24% of the total genome length. Two identical inverted repeat regions (IRA and IRB) of 26,350 bp each collectively constituted 33.22% of the genome, flanking a small single-copy (SSC) region of 18,324 bp, which comprised the remaining 11.55% of the genome.

Fig. 1. Chloroplast genome map of *M. dubia*. The circular map displays the complete cp genome of *M. dubia* (158,664 bp), which exhibits the typical angiosperm quadripartite structure comprising a large single-copy region (LSC), small single-copy region (SSC), and two inverted repeats (IRA and IRB). Genes inside and outside the circle are transcribed in the clockwise and counterclockwise directions, respectively. Colors denote functional categories, including photosystems I and II, cytochrome b/f complex, ATP synthase, NADPH dehydrogenase, ribosomal proteins, tRNAs, rRNAs, and other essential genes. The central photograph shows the *M. dubia* fruits.

Nucleotide composition analysis of the *M. dubia* cp genome revealed a heterogeneous GC content distribution across both structural regions and functional gene categories (Supplementary Figs. 1–2). The overall GC content of the 158,664 bp genome was 37.1%. Regional analysis identified a compositional hierarchy among structural partitions, with inverted repeats (IRA/IRB) exhibiting the highest GC content (42.7%), followed by the large single-copy region (35.0%) and the small single-copy

region (30.9%). Among the functional gene categories, rRNA genes displayed the highest GC content (53.2%), followed by tRNA genes (50.9%), protein-coding genes (PCGs = 37.5%), and intergenic regions (32.4%).

Functional annotation of the cp genome

Comprehensive annotation of the *M. dubia* cp genome identified 135 genomic features distributed across three functional gene categories and four structural regions (Fig. 2, Table 1). PCGs constituted the largest feature class, with 84 sequences representing 60.21% of all annotated elements, followed by 37 tRNA genes, eight rRNA genes (four types each duplicated in both inverted repeats), and five, pseudogenes. Transcriptional orientation analysis revealed category-specific strand asymmetry: 54 PCGs (64%) were oriented on the reverse strand versus 30 (36%) on the forward strand, whereas tRNA genes showed a more balanced distribution (20 reverse, 17 forward), and rRNA genes were equally distributed between strands (four per strand). Regionally, the LSC contained the greatest number of PCGs, the inverted repeats showed elevated rRNA and tRNA densities and the most compact gene arrangement per unit length, and intergenic sequences were the most prevalent in the single-copy regions.

Fig. 2. Structural organization and gene distribution analysis of the cp genome of *M. dubia*. (A) Proportional representation of the four structural regions: LSC, SSC, IRA, and IRB. (B) Genome coverage by annotated feature class, with PCGs as the largest proportion, followed by intergenic regions, tRNA genes, rRNA genes, and pseudogenes. (C) Gene coverage distribution across the overall genome and in each structural compartment. (D) Gene density per unit length across the regions. Color coding is consistent throughout all panels: intergenic regions (gray), pseudogenes (blue), rRNA genes (red), tRNA genes (orange), and PCGs (green).

Gene size analysis revealed pronounced size heterogeneity within each functional category, with no statistically significant differences among the genomic regions (Supplementary Table 1 and Supplementary Fig. 3). PCGs ranged from 90 to 6,861 bp (mean ± SD: 1,062.97 ± 1,304.19 bp), with a strongly leptokurtic and right-skewed distribution (skewness = 2.72; kurtosis = 8.55). The tRNA and rRNA genes averaged 289 ± 498.85 bp (range: 71–2,598 bp) and 1,132 ± 1,199.48 bp (range: 103–2,813 bp), respectively. Although PCGs in the inverted repeat regions were numerically larger (IRA: 1,823.29 ± 2,329.61 bp; IRB: 1,780.00 ± 2,347.88 bp) than those in the single-copy regions (LSC: 844.46 ± 847.85 bp; SSC: 1,311.92 ± 1,543.52 bp), Kruskal–Wallis tests confirmed no significant regional differences for any category (PCGs: $H = 4.028$, $p = 0.258$; tRNAs: $H = 0.421$, $p = 0.936$; rRNAs: $H = 0$, $p = 1.0$). The coefficients of variation were high across all categories (PCGs: 122.7%; tRNAs: 172.6%; rRNAs: 106.0%).

The annotated gene set encoded a complete photosynthetic apparatus, autonomous transcription and translation machinery, and a suite of essential metabolic functions (Table 1). The photosynthetic apparatus comprises five Photosystem I subunit genes (*psaA*–*psaJ*), 14 Photosystem II subunit genes (*psbA*–*psbZ*), six cytochrome b/f complex genes (*petA*–*petN*), six ATP synthase subunit genes (*atpA*–*atpI*), the large subunit of RuBisCO (*rbcL*), and 11 NADH dehydrogenase genes (*ndhA*–*ndhK*). The transcription and translation machinery includes four RNA polymerase subunit genes (*rpoA*, *rpoB*, *rpoC1*, and *rpoC2*), 37 tRNA genes, four rRNA species duplicated in inverted repeats, 11 large ribosomal protein genes (*rpl2*–*rpl36*), and 14 small ribosomal protein genes (*rps2*–*rps19*). Additional functional genes encoded maturase K (*matK*), fatty acid biosynthesis (*accD*), proteolysis (*clpP*), carbon metabolism (*cemA*), and chloroplast import (*ycf1* and *ycf2*). Five pseudogenes were identified: *infA*, two copies of *ycf15*, and two copies of *ycf68*.

Table 1. Comprehensive gene inventory and functional classification of the cp genome of *M. dubia*

Category	Gene group (number)	Gene name
Photosynthesis	Subunits of photosystem I (5)	*psaA, psaB, psaC, psaI, psaJ*
	Subunits of photosystem II (15)	*psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbT, psbZ*

	Subunits of cytochrome b/f complex (6)	*petA, petB*[a], *petD*[a], *petG, petL, petN*
	Subunits of ATP synthase (6)	*atpA, atpB, atpE, atpF*[a], *atpH, atpI*
	Subunits of NADH dehydrogenase (11)	*ndhA*[a], *ndhB*[a,c], *ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*
	Large subunit of rubisco (1)	*rbcL*
Self-replication	Ribosomal RNAs (8)	*rRNA16*[c], *rRNA23*[c], *rRNA4.5*[c], *rRNA5*[c]
	Transfer RNAs (37)	*trnA-UGC*[a,c], *trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnfM-CAU*[e], *trnG-GCC, trnG-UCC*[a], *trnH-GUG, trnI-GAU*[a,c], *trnI-CAU*[c], *trnK-UUU*[a], *trnL-CAA*[c], *trnL-UAA*[a], *trnL-UAG, trnM-CAU, trnN-GUU*[c], *trnP-UGG, trnQ-UUG, trnR-ACG*[c], *trnR-UCU, trnS-GCU, trnS-GGA, trnS-UGA, trnT-GGU, trnT-UGU, trnV-GAC*[c], *trnV-UAC*[a], *trnW-CCA, trnY-GUA*
	Large subunit of ribosome (11)	*rpl2*[a,c], *rpl14, rpl16*[a], *rpl20, rpl22, rpl23*[c], *rpl32, rpl33, rpl36*
	Small subunit of ribosome (14)	*rps2, rps3, rps4, rps7*[c], *rps8, rps11, rps12*[a,d], *rps14, rps15, rps16*[a], *rps18, rps19*
	DNA dependent RNA polymerase (4)	*rpoA, rpoB, rpoC1*[a], *rpoC2*
Other essential functions	Acetyl-CoA carboxylase carboxyl transferase beta subunit (1)	*accD*
	Cytochrome c heme attachment protein (1)	*ccsA*
	Chloroplast envelope membrane protein (1)	*cemA*
	ATP-dependent Clp protease proteolytic subunit (1)	*clpP1*[b]
	Maturase K, RNA splicing and processing (1)	*matK*
	Chaperones for assembly of the photosystem I complex (2)	*ycf3*[b], *ycf4*
	Protein import into chloroplasts (2)	*ycf1*[f], *ycf2*[c]

	Photosystem biogenesis factor 1 (1)	*pbf1*
Pseudogenes	Hypothetical chloroplast reading frame 68 (2)	*ycf68*[c g]
	Translation initiation factor IF-1 (1)	*infA*
	Hypothetical chloroplast reading frame 15 (2)	*ycf15*[c]

[a] Gene containing a single intron.
[b] Genes containing two introns.
[c] Two gene copies in the IRs.
[d] Trans-splicing gene divided into two independent transcription units. Exon 1 is located in the LSC region and copies of the exons 2 and 3 are located in the IRA and IRB regions
[e] Duplicated gene in LSC region
[f] Two gene copies, a short one located in the IRB region, and the larger one partially located in the IRA region
[g] Located in the intron of the gene trnI-GAU

Synonymous codon usage bias analysis

Analysis of codon usage patterns in the cp genome of *M. dubia* identified 22,746 codons from 78 PCGs sequences (Fig. 3 and Supplementary Table 2). Nucleotide composition analysis showed an overall AT content of 62.50% and a GC content of 37.50%. Relative synonymous codon usage (RSCU) analysis identified 30 codons with RSCU values exceeding 1.0 and 32 codons with RSCU values below 1.0, whereas two codons exhibited RSCU values of 1.0. Among codons with RSCU > 1.0, the highest values were observed for TTA (leucine, RSCU = 2.02), GCT (alanine, RSCU = 1.82), and AGA (arginine, RSCU = 1.79).

Fig. 3. Relative synonymous codon usage (RSCU) analysis of PCGs in the cp genome of *M. dubia*. The radial plot displays the RSCU values for all 64 codons; the red dashed circle indicates RSCU = 1.0 (no bias). Bars beyond, within, or at this line denote preferred (RSCU > 1.0; 30 codons), disfavored (RSCU < 1.0; 32 codons), or unbiased (RSCU = 1.0; 2 codons) usage, respectively. The numerical values indicate the individual RSCU scores. Amino acids (single-letter codes) are color-coded according to physicochemical group: nonpolar/hydrophobic (dark teal), nonpolar/aromatic (purple), polar/aromatic (violet), polar/uncharged (green), polar/charged (red), and stop codons (black).

Hierarchical clustering of RSCU values across the 23 Myrtales accessions revealed a consistent genome-wide AT codon preference and three phylogenetically structured accession groups (Fig. 4). The column dendrogram partitioned all 64 codons into two principal branches: a larger branch of AT-ending preferred codons, including TTA (Leu), AGA (Arg), TCT (Ser), and GCT (Ala), each displaying elevated Z-scores (RSCU > 1.0) across all accession rows, and a smaller branch of GC-ending disfavored counterparts, including CTG (Leu), CGC (Arg), TCG (Ser), and GCG (Ala) (RSCU < 1.0). The row dendrogram resolved three accession clusters at k = 3 (mean silhouette = 0.453): Cluster 1 (n = 21), comprising the core Myrtaceae and Combretaceae, including *M. dubia*, displayed a homogeneous, low-variance Z-score profile; within this cluster, accessions MT700489 and MT700490 formed a basal pair. Clusters 2 (n = 1; NC_046574) and 3 (n = 1; NC_031385) occupied outgroup positions with quantitatively distinct Z-score amplitudes, most notably across the six-fold degenerate codon families for leucine, arginine, and serine.

Fig. 4. Dual-dendrogram heatmap of relative synonymous codon usage (RSCU) across 23 Myrtales cp genomes. Rows represent 23 accessions, and columns represent 64 synonymous codons, both ordered by independent hierarchical clustering (UPGMA linkage, Euclidean distance; CPCC = 0.984) on Z-score-normalized RSCU values. Red and blue tones indicate preferred (RSCU > 1.0) and disfavored (RSCU < 1.0) codons, respectively (scale: −3 to +3). The row dendrogram resolved three clusters at k = 3 (mean silhouette = 0.453): Cluster 1 (n = 21, core Myrtaceae and Combretaceae), Cluster 2 (n = 1; NC_046574, Lythraceae), and Cluster 3 (n = 1; NC_031385, Onagraceae). *M. dubia* (PX251231) is indicated in bold green type.

Three complementary analyses across 78 PCGs from 23 Myrtales cp genomes indicated that codon usage bias in *M. dubia* plastid genes is shaped by both mutational pressure and translational selection (Supplementary Figs. 4–5). Wright's ENC plot showed a dataset mean ENC of 46.09 ± 6.65, with 87.2% of *M. dubia* PCGs (68/78) falling below the theoretical expected ENC curve (mean deviation: −5.91 ENC units). Neutrality plot regression slopes were near-zero for all accessions combined (slope = 0.015, p = 0.594) and *M. dubia* alone (slope = −0.0004, p = 0.998). PR2 bias plots revealed consistent strand-

specific asymmetry across all 23 accessions, with T > A and G > C at the third codon positions [A3s/(A3s + T3s) = 0.455; G3s/(G3s + C3s) = 0.538]. Multi-index biclustering across eight gene-level indices reproducibly recovered three accession clusters (k = 3; CCC range: 0.920–0.991), with *M. dubia* consistently assigned to cluster 1 (n = 21, core Myrtaceae–Onagraceae). The GRAVY heatmap separated membrane-associated subunits (mean GRAVY = +0.542) from soluble gene products (mean GRAVY = −0.384; ΔGRAVY = 0.926) in the GRAVY index.

Codon coding and decoding mechanisms

The assignment of all 61 sense codons to five mechanistic decoding categories revealed a hierarchically structured and statistically consistent decoding landscape across all 23 Myrtales cp genomes (Supplementary Table 3; Supplementary Figs. 6–9). The 30 plastid-encoded tRNA species covered all standard amino acids, with codons distributed as follows: Category I, standard Watson–Crick/G_{34} wobble (44 codons; 76.72 ± 7.40%); Category II, U_{34} superwobble (12 codons; 15.85 ± 4.97%); Category III, $A_{34}\rightarrow I_{34}$ inosine decoding via *trnR*-ACG (3 codons; 3.77 ± 2.53%); Category IV, $C_{34}\rightarrow k^2C_{34}$ lysidine modification in *trnI*-CAU (1 codon; 2.78 ± 1.78%); and Category V, two-out-of-three decoding of CGG (1 codon; 0.33 ± 0.57%). A Kruskal–Wallis test across all 1,784 gene observations yielded H = 7,817.86 (df = 4, p < 0.0001), with all ten pairwise Dunn post-hoc comparisons significant after Bonferroni correction (p < 0.0001). Family-level tests remained highly significant (Myrtaceae: H = 6,476.79; Combretaceae: H = 671.96; Lythraceae: H = 337.24; Onagraceae: H = 329.17; all p < 0.0001), and Mann–Whitney U tests detected no significant differences between *M. dubia* and the remaining Myrtaceae accessions for any category.

These category-level differences in proportional distribution were further reflected in the hierarchical biclustering analysis, which resolved two codon-mechanism supergroups and three taxonomically coherent accession clusters across the 23-taxon dataset (Fig. 5). The column dendrogram partitioned the five categories into a major-mechanism clade (Categories I and II) and a minor-mechanism clade (Categories III, IV, and V), with Categories III and IV as sister clades and Category V as the earliest-diverging branch. The accession dendrogram resolved three clusters at k = 3 (silhouette = 0.465): Cluster 1 (n = 4, non-Myrtaceae) showed the highest Category V and lowest Category I proportions; Cluster 2 (n = 3, including *M. dubia*, *Backhousia citriodora*, and *Myrtus communis*) occupied an intermediate position with moderately elevated Category III; and Cluster 3 (n = 16, core Myrtaceae) displayed the highest Category I and lowest Category V proportions. At the functional complex level, all ten complexes showed highly significant inter-category differences (H = 107.37–1,400.78; p < 0.0001). In *M. dubia*, Category I was highest in other/unknown complex genes (82.8%) and lowest in ribosomal subunit genes (LSU: 76.2%; SSU: 76.6%), whereas Category V reached its maximum in RNA polymerase genes (0.69%), exceeding the genome-wide mean (0.37%).

Fig. 5. Hierarchical bi-dendrogram of codon decoding category proportions across 23 Myrtales cp genomes. The heatmap cell color represents the row-wise z-score (diverging scale: dark red = above mean; dark blue = below mean) of the mean gene-level proportions (%) across 78 protein-coding genes; cell text shows actual values. Rows (accessions) and columns (categories) were clustered independently using the Canberra distance and UPGMA linkage (CPCC = 0.9370). The family strip encodes taxonomic affiliation (blue = Myrtaceae, green = Combretaceae, orange = Onagraceae, and purple = Lythraceae). Categories: Cat. I, Watson–Crick/G_{34} wobble; Cat. II, superwobble U_{34}; Cat. III, inosine wobble I_{34}; Cat. IV, lysidine/ava$^2C_{34}$; Cat. V, two-out-of-three.

Simple sequence repeats and long dispersed repeats

SSR and long dispersed repeat analyses identified 1,621 SSRs and 10,300 repeat pairs across the 23 Myrtales cp genomes (Supplementary Table 4; Supplementary Figs. 10–11). SSR counts ranged from 31 (*Ludwigia octovalvis*) to 103 (*Terminalia catappa*), with a mean of 70.5 ± 14.9 per genome; within Myrtaceae, the mean was 70.1 ± 10.7. Mononucleotide repeats dominated (1,482/1,621; 91.4%), predominantly the adenine motif (1,475 occurrences), and 90.1% of SSRs were mapped to intergenic spacers, primarily in the LSC (78.2%). *M. dubia* harbored 69 SSRs, 63 of which were A-type mononucleotides concentrated in the LSC and intergenic spacers. Among the long dispersed repeats, palindromic pairs predominated (9,909/10,300; 96.2%), with 93.4% localized to the IRA and IRB regions. *M. dubia* presented 439 long repeat pairs (428 palindromic), distributed between intergenic spacers (210 pairs) and PCGs regions (194 pairs), with a mean repeat length of 31.5 bp.

Species-level variation in these repeat profiles was resolved by multivariate hierarchical clustering of a 26-feature Z-scored matrix (Fig. 6 and Supplementary Fig. 12). *M. dubia* displayed near-zero Z-scores for SSR_LSC, SSR_mono, SSR_total, and SSR_intergenic features, in contrast to the markedly elevated values of *Terminalia catappa* (MT700489) and *Combretum indicum* (MT700490), and the strongly negative SSR profile of *Ludwigia octovalvis* (NC_031385). The LR_IRA and LR_IRB Z-scores

for *M. dubia* were close to the Myrtaceae family mean, and the LR_forward and LR_reverse scores were slightly negative, a pattern shared with *Plinia trunciflora* (NC_034801), *Eugenia uniflora* (KY392761), and *Psidium guajava* (PP920098), the nearest neighbors in the row dendrogram. Column clustering consistently separated SSR-derived from LR-derived features, with LR_tRNA, LR_rRNA, LR_IRB, and LR_reverse forming discrete sub-clusters at the rightmost partition of the feature dendrogram.

Fig. 6. Hierarchical clustering heatmap of SSR and long repeat feature profiles across 23 Myrtales cp genomes. Z-score-normalized values of the 26 quantitative features are shown. Rows represent individual genomes clustered by Ward's minimum variance linkage (Euclidean distance); the left colour bar indicates family membership: Myrtaceae (blue), Combretaceae (green), Onagraceae (orange), and Lythraceae (purple). The top color bar distinguishes SSR (blue) from the LR (green) features. Red and blue cells indicate values above and below the dataset mean, respectively (scale: −4 to +4), respectively. *M. dubia* (PX251231) is highlighted in bold green as a focal species. SSR, simple sequence repeat; LR, long dispersed repeat; LSC/SSC, large/small single-copy region; IRA/IRB, inverted repeats A and B; CDS, protein-coding sequence

Genome-wide nucleotide diversity and regional variation

Nucleotide diversity varied considerably across the four structural regions of the cp genomes (Fig. 7 and Supplementary Table 5). The SSC region exhibited the highest mean π (0.173, median = 0.178), followed by the LSC region (mean π = 0.084), IRB (mean π = 0.015), and IRA (mean π = 0.010). Kruskal-Wallis analysis revealed statistically significant differences among all four regions ($H > 1000$, $p < 0.001$); Dunn's post-hoc test with Bonferroni correction identified three distinct groups (CLD: LSC = a, SSC = b, IRA = IRB = c), with IRA and IRB constituting the only non-significant pair (Fig. 7B). A prominent diversity hotspot was detected within the SSC at approximately 117.7 kb, where π reached its maximum recorded value of 0.240 across all 3,184 windows analyzed. Frequency distributions indicated that LSC windows were broadly spread around the regional mean, whereas windows falling within both IR regions were strongly concentrated toward low π values, producing markedly right-skewed distributions.

Nucleotide diversity also differed significantly among the four genomic functional categories examined (Kruskal–Wallis test, $p < 0.001$). Intergenic regions displayed the highest mean π (0.096, median = 0.099), followed by PCGs (mean π = 0.071, median = 0.062), tRNA genes (mean π = 0.047, median = 0.009), and rRNA genes (mean π = 0.005, median = 0.005). Dunn's post-hoc test with Bonferroni correction resolved four fully distinct groups, with all pairwise comparisons reaching statistical significance (Fig. 7C). The frequency distributions of π by the genomic region (Fig. 7D) revealed markedly contrasting shapes: LSC windows produced a broad unimodal distribution centered around the regional mean, SSC windows were shifted substantially toward higher π values and displayed the widest spread, while IRA and IRB windows were strongly concentrated near zero, generating pronounced right-skewed distributions consistent with IR-mediated sequence homogenization. The corresponding distributions by functional category (Fig. 7E) showed that rRNA windows formed a narrow, sharply defined cluster near zero, whereas intergenic and PCG windows exhibited broader distributions with extended right-side tails, reaching the highest π values in the dataset. The tRNA category displayed a notably bimodal distribution, with a primary peak near zero and a secondary mode at intermediate π values, reflecting the heterogeneous conservation of tRNA genes across the structural regions.

Fig. 7. Nucleotide diversity (π) across the cp genome of 23 Myrtales taxa (500 bp sliding window, 50 bp steps). (A) Diversity profile along the full genome (~159.6 kb); faint traces show raw π values, and bold lines show Gaussian-smoothed trends. Background shading denotes the four structural regions (LSC, blue; SSC, orange; IRA, green; IRB, purple); the dashed line indicates the global mean π (0.0715); the annotated hotspot marks the highest-diversity window (π = 0.240, ~117.7 kb, SSC). (B, C) Box plots of π by genomic region (B) and category (C); boxes represent IQR, lines indicate medians, and whiskers extend to 1.5× IQR. Compact Letter Display (CLD) above each box summarizes pairwise differences from Dunn's post-hoc test with Bonferroni correction after Kruskal-Wallis (α = 0.05); shared letters indicate non-significant differences. (D, E) Overlapping density histograms of π by genomic region (D) and category (E); the dashed vertical line marks the global mean π, and the sample sizes are indicated in the legend.

Evolutionary rates and selection pressures in PCGs (SWAN analysis)

SWAN six-metric analysis of 78 PCGs across 23 Myrtales cp genomes revealed heterogeneous evolutionary rates, with Myrtaceae consistently lower than outgroup taxa across five of the six metrics (Supplementary Table 6). Dataset-level mean values were: H = 0.003616 (SD = 0.001738), Pvar = 0.024431 (SD = 0.011510), π = 0.025233 (SD = 0.012988), Ka = 0.010273 (SD = 0.010237), Ks = 0.309108 (SD = 0.153659), and ω = 0.050338 (SD = 0.046428). Myrtaceae presented lower mean H (0.00261 vs. 0.00832), Pvar (0.01911 vs. 0.04942), Ka (0.00819 vs. 0.02003), and Ks (0.21938 vs. 0.72996) than outgroup taxa, whereas ω was reversed (Myrtaceae: 0.05272 vs. 0.03887). The most

variable PCGs by H and π were *rpl22*, *ycf1*, *matK*, *ndhF*, and *ndhD*; nine PCGs exceeded mean ω > 0.1, led by *rpl22* (0.2117) and *psaI* (0.1711); no PCG exceeded ω = 1.0. *M. dubia* per-gene means were H = 0.002314, Ka = 0.007674, Ks = 0.205742, and ω = 0.054598, with the highest individual ω values in *rpl22* (0.2495), *rps18* (0.1958), and *rpl20* (0.1803).

The taxonomic and functional structuring identified across individual metrics was comprehensively resolved using a combined six-metric hierarchical clustering heatmap (Fig. 8). The analysis integrated the standardized 468-feature matrix (23 taxa × 78 PCGs × 6 metrics) for species ordering (cophenetic r = 0.9636) and the 78 × 6 mean-metric z-score profile for gene ordering (cophenetic r = 0.7949) analysis. The species dendrogram separated all 19 Myrtaceae accessions from the outgroup taxa (Lythraceae, Onagraceae, and Combretaceae), the first of which formed a high-z-score cluster. Within Myrtaceae, a Myrteae sub-clade comprising *M. dubia*, *A. sellowiana*, *P. guajava*, *P. dioica*, *P. trunciflora*, *P. aureana*, *E. uniflora*, and *R. tomentosa* was separated from a second sub-cluster encompassing the Eucalypteae and allied genera. The gene dendrogram resolved a low-to-moderate z-score group dominated by photosystem, ATP synthase, and cytochrome b6/f genes, and a high-z-score group enriched for ribosomal (*rpl22*, *rps18*), NADH dehydrogenase (*ndhD*, *ndhF*), and other/unknown function genes (*ycf1*, *matK*, *accD*). *M. dubia* displayed predominantly low z-scores across most PCGs, with discrete elevated values at the high-ω loci, *rpl22*, *ycf1*, and *psaI*.

Fig. 8. Hierarchical clustering dendrogram-heatmap of the combined six-metric SWAN analysis across 78 PCGs of 23 Myrtales cp genomes. Species (rows) were ordered using the standardized 468-feature Euclidean distance matrix with average linkage (cophenetic r = 0.9636); genes (columns) were ordered using the 78 × 6 mean-metric z-score profile with average linkage (cophenetic r = 0.7949). Cell colors encode the mean z-score across all six metrics (H, Pvar, π, Ka, Ks, and ω) on a light-blue to navy-blue scale. The vertical strip encodes the family (Myrtaceae, blue; Combretaceae, green; Onagraceae, orange; Lythraceae, purple); the horizontal strip encodes the functional gene complex (see legend). SWAN parameters: window = 3 nt, step = 1 nt; Ka and Ks were calculated using the Nei–Gojobori method with Jukes–Cantor correction.

Comparative genomics

Inverted repeat boundary analysis

A comparative analysis of the four IR junction sites across 23 Myrtales cp genomes revealed a conserved quadripartite genome organization in all accessions (Fig. 9). The total genome sizes ranged from 156,129 bp (*Rhodomyrtus tomentosa*, NC_043848) to 160,326 bp (*Angophora costata*, NC_022412), with *M. dubia* (PX251231) measuring 158,664 bp. The LSC, IRB/IRA, and SSC regions of *M. dubia* spanned 87,640, 26,350, and 18,324 bp, respectively; the two IR copies were of equal length in all 23 accessions. The same set of genes occupied equivalent positions at each junction across all the genomes. In JLB, rps19 extended 0–67 bp into IRB, whereas *rpl2* extended 1,068–1,489 bp into the LSC. At JSB, *ycf1* extended 0–55 bp into the SSC (27 bp in *M. dubia*), and *ndhF* was the first gene within the SSC, proximal to the boundary. At JSA, *ycf1* extended 294–1,695 bp into the IRA (1,408 bp in *M. dubia*), followed by *trnN*-GUU (72 bp) immediately downstream. In JLA, *rpl2* extended 1,068–1,489 bp into the LSC, and *trnH*-GUG (75 bp) was located 0–13 bp from the boundary. This gene arrangement was maintained across Myrtaceae and the Combretaceae, Onagraceae, and Lythraceae representatives included in the comparison.

Fig. 9. Comparative visualization of inverted repeat (IR) boundary junction sites of cp genomes between *M. dubia* and 22 related Myrtales species. Each horizontal row represents one cp genome drawn to a common linear scale; species names and GenBank accession numbers are shown to the left, with *M. dubia* (PX251231) highlighted in teal as the reference. The total genome size (bp) is indicated below each label. Colored rectangles depict the four structural regions: LSC (light blue), IRB/IRA (orange), and SSC (green), with the region sizes labeled accordingly. Vertical lines mark the four junction sites: JLB, JSB, JSA, and JLA. Gene boxes above and below the genome track correspond to the positive and negative strands, respectively; genes spanning a boundary are shown with extension length in parentheses.

Structural organization and gene collinearity in cp genomes

Circos visualization of the 23-taxon dataset revealed a highly uniform gene distribution pattern across all Myrtales cp genomes (Fig. 10A). PCGs, tRNA, and rRNA tiles occupied nearly identical positional intervals across all ideograms. The GC content track displayed a consistent wave-like profile in every genome, with a depression at the SSC and elevated plateaus at the IR boundaries, a pattern conserved across all four families. The gene density histogram showed a bimodal distribution with dense bins in the LSC and IR regions and sparse bins in the SSC region. At the center, 734 inter-genomic PCGs collinearity links connected *M. dubia* to nine other Myrteae species, with the most prominent ribbons directed toward *P. trunciflora* and *Plinia aureana*.

The tribe-level Myrteae comparison resolved these collinearity patterns at a greater visual resolution, confirming complete PCGs synteny among the ten species (Fig. 10B). The ideograms of *M. dubia*, *P. trunciflora*, and *P. aureana* displayed visually indistinguishable PCGs and tRNA tile arrangements, and GC content profiles traced near-identical trajectories across all ten genomes. The 734 collinearity ribbons formed a structured inner web with the heaviest connections linking *M. dubia* to the *Plinia* group (z = 20), followed by progressively lighter ribbons toward the *Myrcia* group (z = 10), the *Pimenta* group (z = 5), *R. tomentosa*, and the *Eugenia* group. No link-crossing events were detected between the LSC and SSC zones across any species pair.

Fig. 10. Circos-based comparative visualization of Myrtales cp genomes with *M. dubia* as the focal species. A) Circos visualization of 23 Myrtales cp genomes, color-coded by family and tribe. Five concentric tracks show the PCGs positions, tRNA genes, rRNA genes, GC content windows, and gene density. Central ribbons display collinearity links between PCGs of *M. dubia* and Myrtea species. (B) Tribe-level comparison of ten Myrteae cp genomes, with *M. dubia* as the focal species. Ideogram colors indicate clade affiliation: *Plinia*, *Myrcia*, *Pimenta*, *Australasian*, and *Eugenia* groups. Ribbon thickness and color encode PCGs' synteny strength and relative phylogenetic proximity to *M. dubia*.

Patterns of genome-wide sequence conservation

The mVISTA alignment demonstrated a high overall degree of sequence identity across the 23 Myrtales cp genomes relative to *M. dubia*, with conservation patterns structured primarily by functional gene category and genomic region (Fig. 11). Within the LSC region, protein-coding gene blocks corresponding to the photosynthetic apparatus, RNA polymerase, and ribosomal protein genes displayed near-continuous high identity across all Myrtaceae members, whereas intergenic regions constituted the principal source of interspecific divergence. The IRA and IRB regions exhibited the most uniformly conserved profiles across the entire dataset, consistent with the strong selective constraint imposed by the rRNA operon and complement of IR-resident genes, including *ndhB*, *rpl2*, and *ycf2*. Within Myrtaceae, Myrteae species maintained high identity to *M. dubia* throughout the LSC and IR regions, with only localized intergenic gaps corresponding to the hypervariable hotspot intervals previously identified in the nucleotide diversity analysis.

Divergence in sequence identity was most pronounced in the SSC region and among the non-Myrtaceae, reflecting the combined effects of increased phylogenetic distance and reduced functional constraints in intergenic intervals. The SSC displayed intermediate and notably heterogeneous conservation across all taxa, with several conspicuous identity gaps concentrated in the intergenic regions flanking *ycf1* and *ndhF*. The four non-Myrtaceae species, *T. catappa*, *C. indicum*, *L. octovalvis*, and *C. hyssopifolia*, showed substantially reduced alignment continuity across both coding and non-coding intervals compared to Myrtaceae members, with the greatest discontinuities observed in *L. octovalvis* and *C. hyssopifolia*, consistent with their placement in distantly related families. The Eucalypteae representatives (*S. quadrifida*, *A. ternata*, *E. grandis*, *C. eximia*, and *A. costata*) presented a slightly more heterogeneous identity profile than the Myrteae species, particularly across the SSC, reflecting their phylogenetic distance from *M. dubia*.

Fig. 11. mVISTA pairwise sequence identity alignment of 23 Myrtales cp genomes relative to *M. dubia* as a reference. Alignments were generated using the NUCmer algorithm with a sliding window of 200 bp and a minimum identity threshold of 50%. The reference genome was annotated at the top of the figure, with genomic regions delineated as LSC (sky blue), IRA and IRB (orange), and SSC (green). Gene features are classified as PCGs (dark blue), tRNA and rRNA genes (dark green), and intergenic regions (orange). Each horizontal panel represents the pairwise sequence identity profile of one comparison genome against *M. dubia* across the full reference coordinates (x-axis, kb). The panel order follows a phylogenetic arrangement from closely related Myrteae species (top) to non-Myrtaceae (bottom). Color coding within the identity profiles follows the annotation legend shown at the bottom of the figure.

Phylogenetic analysis

Maximum likelihood phylogenetic reconstruction based on cp genome sequences resolved *M. dubia* within a well-supported monophyletic Myrtales clade (bootstrap = 100%), clearly distinct from the six additional angiosperm orders represented in the analysis. The Myrtales clade comprised 23 accessions spanning four families: Myrtaceae, Combretaceae, Onagraceae, and Lythraceae (Fig. 12). The GTR+F+R4 substitution model, selected as the best fit based on the Bayesian Information Criterion, produced a consensus tree with branch lengths proportional to the evolutionary distance (scale bar: 0.01 substitutions per site). The broader topology recovered Lamiales, Ericales, Sapindales, Malvales, Brassicales, and Rosales as well-supported monophyletic groups (bootstrap = 99–100%), with *Medicago truncatula* (Fabales) as the outgroup.

Within the Myrtales clade, Myrtaceae was recovered as monophyletic with maximum bootstrap support (100%), and *Myrtus communis* occupied the most basal position within the family. The remaining 18

Myrtaceae accessions were resolved into three well-supported subclades. The first comprised the core Myrteae, in which *M. dubia* formed a strongly supported sister relationship with *P. trunciflora* (bootstrap = 100%), which was progressively nested within a larger assemblage that included *M. amethystina* (bootstrap = 100%), *P. guajava* (bootstrap = 99%), *P. dioica*, and *A. sellowiana* (bootstrap = 99%). A sister Myrteae sub-clade (bootstrap = 87%) united *P. aureana*, *E. uniflora*, *L. confertus*, and *B. frutescens*. *R. tomentosa* was recovered as an independent lineage within Myrtaceae and did not cluster with either Myrteae sub-clade. The second subclade (bootstrap = 100%) comprised the Eucalypteae representatives *C. eximia*, *A. costata*, and *E. grandis* (bootstrap = 100%), together with *S. quadrifida* and *A. ternata* (bootstrap = 100%). The third subclade united *S. acuminatissimum* and *B. citriodora* (bootstrap = 100%).

Fig. 12. Maximum-likelihood phylogenetic tree of *M. dubia* and related species based on complete cp genome sequences. The phylogenetic tree was reconstructed using the GTR+F+R4 model selected by IQ-Tree based on the Bayesian Information Criterion. Sequences were aligned using MAFFT and cleaned with BMGE via NGPhylogeny.fr. Bootstrap support values (1,000 replicates) are shown numerically (%) for each node. Background shading denotes taxonomic orders: Myrtales (green), Lamiales (yellow), Ericales (orange), Sapindales (pink), Malvales (light pink), Brassicales (lavender), and Rosales (red). Within Myrtales, right-hand bars indicate family membership: Myrtaceae (blue), Combretaceae (green), Onagraceae (orange), and Lythraceae (purple*). Medicago truncatula* (Fabales) was used as an outgroup. The scale bar represents 0.01 substitutions per site.

Discussion

Chloroplast Genome Organization and Evolutionary Patterns

The quadripartite architecture of the *M. dubia* cp genome exemplifies the remarkable structural conservation that has characterized plastid evolution across angiosperms for more than 100 million years [29]. This organizational pattern, with inverted repeats separating large and small single-copy regions, represents an evolutionary solution to the competing demands of genomic stability and adaptive flexibility [30]. The persistence of this architecture across diverse angiosperm lineages, despite substantial variation in genome size and gene content, suggests that the quadripartite structure provides fundamental advantages for chloroplast function and inheritance [1,31]. These advantages likely include enhanced genome stability through homologous recombination between IR copies, maintenance of stoichiometric balance in gene dosage for rRNA operons, and facilitation of heteroplasmic sorting during organellar transmission [32–35].

The heterogeneous GC content distribution across genomic compartments, with the inverted repeats exhibiting the highest GC content, followed by the LSC and SSC, and with rRNA genes reaching the highest values among gene categories and intergenic regions the lowest, reflects the intricate interplay between mutational biases, selective constraints, and structural requirements that shape cp genome evolution [34,36,37]. The elevated GC content in IR regions and RNA genes can be mechanistically attributed to several factors: stronger purifying selection against AT-biased mutations in functionally critical sequences, structural requirements for stable secondary structures in ribosomal RNAs, where GC base pairing provides superior thermodynamic stability, and potentially localized differences in DNA repair efficiency or mutational pressures across genomic regions [38–40]. This compositional heterogeneity is not merely a neutral byproduct of genome organization but represents an optimized solution that balances the conflicting demands of protein-coding requirements, RNA structural stability, and overall genomic maintenance [41–43]. The observation that intergenic spacer regions show reduced GC content supports the hypothesis that relaxed functional constraints in non-coding sequences permit compositional drift toward AT-richness, the ancestral state for organellar genomes [44].

The presence of pseudogenes, which represent genes in various stages of functional decay, provides a molecular window into the ongoing evolutionary processes that shape cp genome content across plant lineages [45,46]. The identification of five pseudogenes (*infA*, two copies of *ycf15*, and two copies of *ycf68*) in *M. dubia* is consistent with broader patterns of cp genome streamlining observed throughout angiosperms, where genes with redundant functions or those whose products can be imported from the nucleus become dispensable and gradually degenerate [47,48]. This process of gene loss and functional transfer from organellar to nuclear genomes represents a major evolutionary trend in plant evolution, driven by the advantages of centralized genetic control, reduced mutation rates in nuclear DNA, and elimination of conflicts between cytoplasmic and nuclear genetic interests [49,50]. Understanding these patterns of gene retention and loss is essential for reconstructing the evolutionary trajectories of cp genomes and predicting future evolutionary changes in organellar genome architecture [51,52].

Codon Usage Patterns

The pronounced AT compositional bias in chloroplast protein-coding genes reflects a fundamental evolutionary trajectory shaped by multiple selective pressures on organellar translation systems [53,54].

While initially appearing maladaptive, given that GC-rich sequences typically confer greater thermodynamic stability, this AT bias can be explained by several complementary mechanisms [55,56]. The reduced effective population size of organellar genomes compared to that of nuclear genomes weakens selection efficiency, permitting mutational biases toward AT to predominate over weak selection for optimal codon usage [57]. Additionally, the unique biochemical environment of chloroplasts, with high ATP/ADP ratios and alkaline stroma pH, may impose different selective pressures on translation compared to the cytoplasm, potentially favoring AT-rich codons that match the available tRNA pool and optimize translation kinetics under chloroplast-specific conditions [40,58,59].

The systematic preference for codons ending in A or T represents an evolutionary compromise between the mutational landscape and functional requirements of chloroplast translation. This pattern emerges from the interplay of directional mutation pressure toward AT nucleotides, selection for translation efficiency matching the chloroplast tRNA complement, and constraints imposed by the reduced chloroplast tRNA gene repertoire [60–62]. ENC–GC3s analysis corroborated this interpretation, with most *M. dubia* PCGs falling below the theoretically expected ENC curve (Supplementary Fig. 4), indicating that translational selection contributes to codon bias beyond what is expected from mutational pressure alone. The near-zero slopes of the neutrality plot for both the full dataset and *M. dubia* alone indicate that mutational pressure and translational selection operate approximately equally, with neither force dominating codon evolution in the plastid genomes.

The convergence of codon usage patterns across phylogenetically diverse angiosperm lineages suggests that these selective pressures are sufficiently strong and consistent to override lineage-specific mutational biases in angiosperms. This indicates that codon usage in chloroplasts is actively maintained by selection rather than simply drifting toward mutational equilibrium [63,64]. The identification of specific optimal codons, which are preferentially used in highly expressed genes, demonstrates that chloroplast translation has evolved sophisticated mechanisms for matching codon usage to tRNA availability, maximizing translational speed and accuracy for photosynthetic genes, the expression levels of which critically determine plant fitness [65–67].

Codon coding and decoding mechanisms

The five-category decoding framework established across all 23 Myrtales plastomes reveals a highly structured and evolutionarily constrained translation system that extends well beyond canonical Watson–Crick base pairing [61,68,69]. Category I codons, decoded through standard Watson–Crick interactions and G_{34} wobble, account for the majority of codon usage, a proportion that was statistically indistinguishable between *M. dubia* and other Myrtaceae accessions, reflecting the central importance of canonical decoding mechanisms in maintaining translational fidelity across photosynthetic organisms [68,69]. Category II codons (U_{34} superwobble) illustrate an elegant molecular economy in which a single tRNA species reads all four synonymous codons in a fourfold degenerate box without an isoacceptor partner, a mechanism that effectively compensates for the reduced tRNA gene repertoire characteristic of plastid genomes and has been experimentally validated in *Nicotiana tabacum* transplastomic knockouts [61,70].

The three post-transcriptional modification categories (III, IV, and V), which collectively account for less than 7% of codon usage, represent the most biochemically sophisticated decoding strategies in plant organellar translation. Inosine modification in Category III ($A_{34}\rightarrow I_{34}$ in trnR-ACG), lysidine modification in Category IV (trnI-CAU), and the two-out-of-three mechanism in Category V (CGG) each involve enzymatic anticodon modifications critical for complete codon coverage, despite the markedly reduced tRNA complement [71–75]. The Kruskal–Wallis test confirmed that these five categories are quantitatively distinct features of the plastid decoding landscape, rather than technical classifications without functional significance. Taxon-level clustering, in which non-Myrtaceae families form a discrete cluster with elevated Category V and reduced Category I proportions, likely reflects differences in tRNA gene complement and anticodon sequence divergence between families rather than fundamental differences in decoding strategy. The elevated Category V proportion in the RNA polymerase genes of *M. dubia* is consistent with the higher CGG frequency in transcriptional machinery genes and may reflect a functional coupling between RNA polymerase gene codon composition and the efficiency of plastid transcription [72].

Nucleotide Diversity and Selection Pressures

The substantial variation in evolutionary rates across chloroplast genomic regions reflects the fundamental tension between mutational processes and functional constraints that govern organellar genome evolution [76,77]. The gradient of diversity from single-copy to inverted repeat regions can be mechanistically explained by gene conversion operating between IR copies, which homogenizes sequences and effectively reduces the fixation rate of new mutations through continuous sequence

correction [30,32]. This stabilizing mechanism represents a key evolutionary innovation in plastid genome architecture, as it maintains sequence fidelity in genes critical for photosynthetic function while permitting greater evolutionary flexibility in single-copy regions where recombinational repair is absent [78]. The conservation of this architectural pattern across angiosperms suggests that the quadripartite structure arose early in land plant evolution and has been maintained because it provides an optimal balance between genomic stability and evolutionary adaptability [79,80].

The identification of high-diversity loci holds significant practical value for understanding the population structure and evolutionary history within Myrtaceae, particularly for taxa such as *M. dubia* that occupy fragmented Amazonian habitats subject to anthropogenic disturbance [81–84]. Variable chloroplast regions serve as powerful tools for reconstructing recent demographic events, including population bottlenecks, range expansions, and gene flow patterns that may have resulted from Quaternary climate fluctuations or more recent habitat fragmentation [85,86]. The development of these molecular markers is particularly urgent for Amazonian species, where rapid deforestation and climate change threaten genetic diversity before it can be adequately characterized [84,87]. Moreover, population-level chloroplast variation can inform conservation prioritization by identifying genetically distinct populations that represent unique evolutionary lineages deserving protection as evolutionarily significant units [35,88].

The contrasting selection regimes evident from the Ka/Ks ratio patterns illuminate how natural selection differentially maintains chloroplast gene function across evolutionary timescales [89,90]. The low dataset-level mean ω, combined with the absence of any PCG exceeding ω = 1.0, confirms that all 78 examined protein-coding genes are evolving under net purifying selection across the 23 Myrtales accessions. Genes exhibiting Ka/Ks ratios near zero experience extremely strong purifying selection, reflecting the severe fitness consequences of even conservative amino acid substitutions in core metabolic pathways, where protein function is finely tuned and co-evolved with multiple interaction partners [3,91]. In contrast, the small set of PCGs exceeding mean ω > 0.1, (led by rpl22 and psaI), together with *ycf1*, *matK*, and *ndhF,* displayed significantly relaxed purifying selection relative to the dataset mean, suggesting lineage-specific functional refinements or reduced functional constraint in these loci [92–94]. The elevated ω values among *M. dubia* ribosomal protein genes specifically (*rpl22*, *rps18*, and *rpl20*), may reflect ongoing structural co-evolution between plastid-encoded ribosomal proteins and their nuclear-encoded interaction partners rather than adaptive evolution [94]. Importantly, the absence of any gene with ω ≥ 1.0 across the entire dataset indicates that positive selection is not detectable at the gene level with the current 23-taxon sampling, and interpretations of accelerated evolution in these loci should be framed as relaxed purifying selection rather than positive selection.

Repetitive Element Distribution and Genome Plasticity

The distribution patterns of simple sequence repeats across cp genomes reflect the dynamic balance between mutation-driven repeat generation and selection-mediated repeat suppression in different genomic contexts [95]. The concentration of SSRs in single-copy regions rather than inverted repeats can be mechanistically explained by the homogenizing effects of gene conversion between IR copies, which actively eliminates polymorphic repeats and maintains sequence identity [32]. This differential repeat accumulation across genomic compartments has important evolutionary consequences: SSR-rich single-copy regions may serve as reservoirs of genetic variation and potential sources of adaptive genomic rearrangements, whereas SSR-depleted IR regions maintain stability essential for ribosomal RNA genes and other dosage-sensitive loci [96]. The predominance of mononucleotide repeats, particularly poly A/T tracts, reflects both the underlying AT-biased mutational spectrum of cp genomes and the increased slippage propensity of homopolymer sequences during DNA replication [97].

The functional significance of repetitive elements extends beyond their role as neutral evolutionary markers to encompass their potential regulatory functions and contributions to genome evolution. SSRs located in gene promoters or untranslated regions may modulate gene expression through length-dependent effects on DNA structure, transcription factor binding, or RNA stability, although the extent of such regulatory roles in chloroplasts remains incompletely characterized [98]. More dramatically, larger repetitive elements can facilitate illegitimate recombination events, leading to gene duplications, deletions, or inversions, which are the raw materials for major evolutionary innovations in cp genome architecture [99]. The observed regional heterogeneity in repeat distribution may therefore reflect historical selection against recombinogenic repeats in functionally critical genome regions, while tolerating or even favoring repeat accumulation in regions where genomic rearrangements are neutral or occasionally beneficial [100].

Long dispersed repeats exhibited a strikingly different genomic distribution from SSRs, with palindromic pairs concentrated predominantly in the IRA and IRB regions of the genome. This IR-enrichment of

palindromic repeats is mechanistically consistent with the inverted orientation of these regions and suggests that palindromic repeats may preferentially arise through intra-IR recombination or be tolerated in IR regions where the disruption of coding sequences would be corrected by gene conversion [99,100]. The near-equal distribution of long repeats in *M. dubia* between intergenic spacers and PCGs regions is noteworthy, as repeats located within PCGs could influence protein structure through amino acid repeat domains or affect transcript stability, whereas those in intergenic regions may modulate regulatory signals [98,101]. These repetitive elements may contribute to cp genome plasticity by providing substrates for slippage-mediated length variation, recombination-mediated rearrangements, and potential regulatory innovation through repeat expansion or contraction [102]. Understanding the evolutionary dynamics of these repetitive elements is essential for predicting cp genome stability, interpreting patterns of genomic variation, and harnessing repeat-mediated mechanisms for synthetic biology applications targeting chloroplast transformation [103].

Comparative Chloroplast Genome Architecture and Synteny

A comparative analysis of IR boundary junctions across all 23 Myrtales accessions demonstrated complete conservation of gene positioning at all four junction sites: *rps19* and *rpl2* at JLB; *ycf1* and *ndhF* at JSB; *ycf1* and *trnN-GUU* at JSA; and *rpl2* and *trnH-GUG* at JLA across the four families and deep evolutionary time. This uniformity is remarkable, given that IR boundary fluctuations have been reported as a source of substantial structural diversity in many other angiosperm families, in which the expansion or contraction of the IR into adjacent single-copy regions alters the gene complement and dosage of the inverted repeats [35,37,91]. The conservation of boundary gene identity in Myrtales suggests that strong functional constraints on genes, such as *ycf1* and *ndhF,* at the SSC junctions limit the tolerable range of IR boundary displacement, thereby stabilizing the quadripartite structure across this order.

The complete PCG synteny was confirmed by Circos-based analysis of 734 CDS collinearity links, with no link-crossing events detected between the LSC and SSC zones across any species pair, reinforcing earlier evidence that Myrtaceae cp genomes are among the most structurally stable in angiosperms [104,104,105]. The progressive decrease in ribbon thickness from the *Plinia* to the *Myrcia* and *Eugenia* groups reflects increasing phylogenetic distance while maintaining complete gene order conservation, a pattern consistent with the expectation that intrafamily genomic rearrangements are too infrequent in Myrtaceae to serve as informative phylogenetic characters at the tribal level. This structural conservation also validates the analytical framework of the present study: the use of a common set of 78 PCGs as the basis for all comparative analyses is justified because gene content and order are invariant across the dataset.

The mVISTA sequence identity profiles quantify the divergence gradient across the coding and non-coding compartments at the nucleotide level. The near-continuous high identity observed in IR-resident rRNA and photosynthetic gene blocks across all Myrtaceae members is consistent with the extremely low nucleotide diversity of the IR regions. Both lines of evidence converge on the conclusion that gene conversion between IR copies is the dominant force maintaining sequence fidelity in these compartments. Conversely, the concentration of identity gaps in SSC intergenic regions, particularly flanking *ycf1* and *ndhF*, directly corroborates the diversity hotspot identified at ~117.7 kb in the sliding-window analysis. Together, these two analyses identified the same genomic intervals as the primary targets for population genetic marker development in *M. dubia* and closely related Myrteae species. The substantially reduced alignment continuity of the four non-Myrtaceae outgroups, most pronounced in *L. octovalvis* and *C. hyssopifolia*, was consistent with their greater phylogenetic distances.

Phylogenetic Relationships and Taxonomic Implications

The phylogenetic positioning of *M. dubia* as sister to *Plinia trunciflora* with maximum bootstrap support, nested within the core Myrteae together with *Myrcia amethystina*, *Psidium guajava*, *Pimenta dioica*, and *Acca sellowiana*, provides robust molecular evidence for the close evolutionary relationships within the *Myrciaria–Plinia* complex of Myrtaceae [106,107]. This strongly supported sister relationship raises important questions about generic boundaries within the Myrteae tribe, where paraphyly involving *Myrciaria* and *Plinia* and broader taxonomic instability have been documented [108,109]. The recovery of three well-supported Myrtaceae sub-clades—(1) core Myrteae; (2) Eucalypteae comprising *Corymbia eximia*, *Angophora costata*, and *Eucalyptus grandis* with *Stockwellia quadrifida* and *Allosyncarpia ternata*; and (3) a clade uniting *Syzygium acuminatissimum* and *Backhousia citriodora*, all with bootstrap = 100%, confirms the family-level tribal relationships previously established by molecular systematics [110,111]. This pattern of deep, well-resolved lineage separation suggests that the *Myrciaria-Plinia* lineage represents an ancient component of Myrtaceae diversity that may have originated during the early diversification of the family in the Paleogene [112,113].

The phylogenetic distinctiveness of *M. dubia* has significant biogeographic implications for understanding the diversification of Amazonian Myrtaceae. The sister relationship between *P. trunciflora* and its partial distribution suggests *in situ* diversification within the Amazon Basin, consistent with riverine barrier models and climate-driven habitat fragmentation during the Quaternary [114,115]. This pattern aligns with broader phylogeographic evidence indicating that Amazonian plant diversification has been shaped by complex interactions between landscape evolution, climate oscillations and ecological opportunity [116,117].

From a conservation perspective, the basal phylogenetic position of *M. dubia* within Myrtaceae indicates high phylogenetic distinctiveness, which is increasingly recognized as an important criterion for conservation prioritization [118,119]. Species occupying deep phylogenetic positions often harbor unique evolutionary adaptations and are irreplaceable components of biodiversity. The complete synteny and lack of structural rearrangements observed across Myrtaceae cp genomes reinforce the phylogenetic signal and validate the utility of plastome sequences in resolving systematic relationships in this taxonomically complex family [104,120,121].

Limitations and Future Directions

Several limitations of this study should be acknowledged when interpreting the results. First, the analysis was based on a single individual, limiting insights into intraspecific variation patterns, which would be valuable for population genetic studies. Future research should incorporate multiple individuals across the natural distribution of *M. dubia* to assess cp genome diversity and identify population-specific markers.

Second, although the comparative analysis included 19 Myrtaceae species, expanding this dataset to include additional Myrciaria species and closely related genera would enhance phylogenetic resolution and improve our understanding of the evolutionary relationships within this economically important group. The limited availability of complete cp genomes for Amazonian Myrtaceae species represents a significant constraint that future sequencing efforts should address.

Future research should focus on several key issues. Population-level studies using the molecular markers identified herein could reveal genetic diversity patterns essential for conservation planning, particularly in light of the increasing anthropogenic pressure on Amazonian ecosystems. Comparative transcriptomic analyses could identify functional differences in chloroplast gene expression between *M. dubia* and related species, potentially revealing the molecular basis of the exceptional vitamin C production in this species. Additionally, integrating chloroplast genomic data with nuclear and mitochondrial genome information would provide comprehensive insights into the evolutionary history and genomic architecture of *M. dubia*. Such multi-genome approaches are becoming increasingly important for understanding complex evolutionary processes and developing effective conservation strategies for tropical plant species.

Conclusions

This study presents the first complete cp genome of *M. dubia* and the most comprehensive comparative plastid genomic analysis of the species to date, integrating structural annotation, codon biology, nucleotide diversity, repeat characterization, and phylogenetic reconstruction across 23 Myrtales accessions. The analyses collectively establish that the *M. dubia* cp genome conforms to the conserved quadripartite architecture characteristic of Myrtaceae, with codon usage shaped by the interplay of mutational pressure and translational selection, complete codon coverage achieved through five mechanistic decoding categories, and evolutionary rates governed by pervasive purifying selection across all PCGs. Comparative genomics confirmed complete gene collinearity and conserved IR boundary architecture across 19 Myrtaceae accessions, while phylogenetic reconstruction provided a robust molecular framework positioning *M. dubia* within the Myrteae tribe as sister to *P. trunciflora*. The genomic resources and molecular markers generated herein, particularly the hypervariable SSC intergenic regions identified as priority targets for population-level studies, establish an essential foundation for conservation genetics, germplasm characterization, and breeding programs for this economically and nutritionally important Amazonian species. More broadly, the analytical frameworks demonstrated are directly applicable to other understudied tropical cp genomes, contributing to the growing genomic infrastructure required to document, understand, and conserve the extraordinary plant diversity of the Amazon Basin.

Materials and Methods

Plant material collection and genomic DNA purification

Fresh young leaves of *M. dubia* were collected from the *ex situ* germplasm bank at the Experimental Field “El Dorado” (03°57′17″S, 73°24′55″W, 112 m.a.s.l), located 25.5 km from the Iquitos-Nauta highway in Loreto, Peru.

Total genomic DNA was purified using a modified cetyltrimethylammonium bromide (CTAB) protocol optimized for high-polyphenol tropical plant tissues [122]. Briefly, 200 mg of fresh leaf tissue was ground and incubated in CTAB extraction buffer (2% CTAB, 20 mM EDTA, 100 mM Tris-HCl pH 8.0, 1.4 M NaCl) supplemented with 2% polyvinylpyrrolidone (PVP) and 0.2% β-mercaptoethanol for 60 min at 65°C. DNA was purified using chloroform-isoamyl alcohol extraction and precipitated with isopropanol. DNA purity was assessed spectrophotometrically using a NanoDrop 2000 (Thermo Scientific, Waltham, MA, USA); samples with A_{260}/A_{280} ratios ≥ 1.8 were considered acceptable for downstream processing. DNA concentration was quantified by fluorometry using a Qubit™ 4.0 Fluorometer with the Qubit™ dsDNA High Sensitivity Assay Kit (Thermo Fisher Scientific, Waltham, MA, USA), following the manufacturer's instructions.

Library Preparation and High-Throughput Sequencing

Sequencing libraries were prepared using the Illumina TruSeq DNA PCR-Free Library Prep Kit, following the manufacturer's instructions. Genomic DNA (1 µg) was fragmented using a Covaris S220 ultrasonicator (Covaris, USA) to obtain average fragment sizes of 350 bp. The fragmented DNA was end-repaired, A-tailed, and ligated with adapters. Library quality was assessed using an Agilent Bioanalyzer 2100 (Agilent Technologies, USA), and the library concentration was determined via quantitative PCR using the KAPA Library Quantification Kit. Paired-end sequencing (2 × 150 bp) was performed using an Illumina NovaSeq 6000 platform.

Chloroplast Genome Assembly and Quality Assessment

Raw sequencing reads were quality-filtered using Trimmomatic v0.39 [123] with the following parameters: ILLUMINACLIP and TruSeq3-PE. fa:2:30:10, LEADING:3, TRAILING:3, SLIDINGWINDOW:4:15, MINLEN:36. Chloroplast genome assembly was performed on the Galaxy platform [124] using Novoplasty v4.3.1 [125] and GetOrganelle v1.7.7 [126] with default parameters that were optimized for angiosperm chloroplasts. Assembly quality was evaluated using a series of complementary verification steps performed in Geneious Prime® 2025.2.1 (Biomatters Ltd., Auckland, New Zealand). Quality-filtered reads were remapped to the assembled sequence using the Map to Reference function, and the mean sequencing depth and per-base coverage uniformity were calculated across the full cp genome to confirm adequate and homogeneous read representation. Circularity was verified by inspecting the coverage histogram in circular viewing mode and confirming the presence of reads spanning the start/end junction boundary, where paired-end reads were expected to map continuously across the artificial linear break. The absence of residual terminal overlap in the finalized sequence confirmed that the assembler correctly resolved the circular topology. Structural integrity was further corroborated by confirming the expected quadripartite organization, LSC, IRA, SSC, and IRB, and the canonical positioning of the boundary genes, including *rpl2*, *rps19*, *ycf1*, and *ndhF*, at all four IR junction sites. The final verified assembly was retained in Geneious Prime for subsequent annotation and downstream analysis.

Genome Annotation and Gene Prediction

Initial genome annotation was performed using GeSeq v1.93 [127] with default parameters for cp genomes, utilizing reference sequences from closely related Myrtaceae species belonging to the tribe Myrteae (*Pimenta dioica* [NC_034684.1], *Psidium guajava* [NC_033355.1], *Plinia aureana* [NC_039557.1], and *Rhodomyrtus tomentosa* [NC_043848.1]). Transfer RNA genes were identified and verified using tRNAscan-SE v2.0 [128] with default settings for cp genomes. Ribosomal RNA genes were identified using BLAST searches against chloroplast rRNA databases and manual curation. PCGs were annotated through homology searches and open reading frame analysis with manually verified start and stop codons. All annotations were manually curated using Geneious Prime® 2025.2.1 (Biomatters Ltd., New Zealand) to ensure accuracy and consistency. A circular cp genome map was generated using OGDRAW v1.3.1 [129].

Synonymous Codon Usage Bias Analysis

To perform this analysis, a dataset comprising 23 complete cp genome sequences from representatives of the order Myrtales was retrieved from the NCBI GenBank database. The focal species was *Myrciaria dubia* (Kunth) McVaugh “camu-camu” (accession PX251231), a native Amazonian species. The remaining 22 accessions represented diverse Myrtales families, including Myrtaceae, Melastomataceae, and Onagraceae (Supplementary Table 7). For each genome, a total of 78 PCGs were identified and used as the basis for codon frequency calculations, following the standard annotation framework for plastid genomes.

Relative synonymous codon usage (RSCU) values were computed for all 78 PCGs across the 23 Myrtales accessions using the RSCU subcommand of CPStools v3.0 [130], with methionine (ATG) and tryptophan (TGG) retained in the dataset but excluded from downstream multivariate clustering analyses

because of their invariant RSCU values. For each amino acid and its synonymous codons, RSCU was defined as the observed codon count divided by the expected count under equal use of all synonymous codons [131,132]. Overall RSCU values were obtained by summing codon counts across all genes using the following equation:

$$RSCU_{ij} = \frac{X_{ij}}{\frac{1}{n_j} \sum X_i}$$

where X_{ij} is the observed count of the i□□ codon for amino acid j, and $n_{□}$ is the number of synonymous codons for that amino acid. An RSCU value equal to 1.0 indicates equal usage among synonymous codons, values greater than 1.0 indicate preferential usage, and values less than 1.0 indicate avoidance. Methionine (ATG) and tryptophan (TGG), each encoded by a single codon, were retained in the dataset but excluded from the downstream clustering of codons, as their RSCU values are by definition invariant at 1.0. Because RSCU values are bounded by synonymous family size (maximum = $n_{□}$, from 2 for two-fold to 6 for six-fold degenerate families), raw values are heteroscedastic across families and must be standardized prior to Euclidean distance-based analyses [133]. A global Z-score transformation [z = (RSCU − μ) / σ, computed per codon across all 23 accessions] was applied to place all codons on a common scale. The two invariant single-codon amino acids (Met/ATG; Trp/TGG) were excluded from multivariate analyses.

Hierarchical agglomerative clustering was performed on the 23 × 62 Z-score matrix of the RSCU dataset. Seven linkage–distance combinations were evaluated; the optimal method was selected by maximizing the cophenetic correlation coefficient (CCC), a measure of dendrogram fidelity [134]. UPGMA (average linkage) with Euclidean distance yielded the highest CCC (0.9840) and was adopted for all analyses. Codon clustering (column ordering) used correlation distance on the transposed matrix, which captures co-usage patterns independently of scale [135]. The optimal number of clusters k was determined by maximizing the mean silhouette score [136] across k = 2–7; k = 3 was selected (silhouette = 0.453), consistent with the three major phylogenetic lineages in the dataset. The dual-dendrogram heatmap was generated with seaborn v0.136 in Python 3.11, using a diverging blue–white–red colour scale centered at Z = 0 (range −3.5 to +3.5).

Also, based on the 78 PCGs, a comprehensive codon usage bias analysis was conducted using a custom Python script based on CodonW v1.4.4 [137]. For every gene, was derived a complete codon count profile across all sense codons of the standard genetic code (excluding stop codons). GC content at the first, second, and third codon positions (GC1, GC2, GC3) was computed by counting G and C nucleotides at each respective position and dividing by the total number of sites. GC12 was calculated as the arithmetic mean of GC1 and GC2. The effective number of codons (ENC) was computed using an approximation based on the homozygosity of codon usage within synonymous families, excluding the non-degenerate codons for methionine and tryptophan. The expected ENC under a neutral model driven only by GC3 was calculated using the standard ENC–GC3 relationship [132]:

$$ENC_{exp} = 2 + \frac{29}{GC3^2 + (1 - GC3)^2}$$

A codon adaptation index (CAI)-like measure was estimated using overall RSCU values to define relative codon weights. For each amino acid, codon weights were set as the RSCU value divided by the maximum RSCU among its synonymous codons. For each gene, the CAI proxy was defined as the geometric mean of weights across all codons in the gene [131,138]. Translated amino acid sequences were analyzed with Biopython to estimate the grand average of hydropathicity (GRAVY) and the aromaticity (AROMO) index for each encoded protein.

Parity Rule 2 (PR2) bias at the third codon position was assessed by computing the relative proportions A3/(A3 + T3) and G3/(G3 + C3) for each gene [139]. These ratios were used to construct a PR2-bias plot to visualize deviations from equal use of complementary bases at silent sites. A neutrality plot was constructed by plotting GC12 against GC3 for all genes and fitting a linear regression. The slope and dispersion of points around the regression line provide insight into the relative influence of mutation pressure versus selection on codon usage. The relationship between ENC and GC3 was examined by plotting the observed ENC values against GC3 and overlaying the theoretical ENC–GC3 curve. Genes falling substantially below the expected curve are interpreted as being under stronger selection for codon bias than expected from mutational pressure alone.

Genes were ranked according to the CAI proxy. The upper and lower 10% of genes were taken as proxies for highly and weakly expressed genes, respectively. For each codon, mean RSCU was

computed in the high- and low-CAI subsets, and the difference in RSCU (ΔRSCU) was used to define optimal codons. A codon was considered optimal if ΔRSCU > 0.08 and RSCU > 1 in the high-CAI group. The frequency of optimal codons (FOP) was calculated per gene as the number of optimal codons divided by the total number of synonymous codons (excluding methionine, tryptophan, and stop codons) [140]. A simple codon bias index (CBI-like) was computed as (number of optimal codons − number of non-optimal codons) divided by the total number of synonymous codons [141].

For each of the eight indices described above (ENC, ENC ratio, GC3s, GC12, GRAVY, AROMO, CAI, and CBI), gene-level data were arranged into an accession × gene matrix (23 × 78). Both rows (accessions) and columns (genes) were independently clustered using hierarchical agglomerative clustering with UPGMA linkage and Euclidean distance [134,142]. The quality of the hierarchical clustering solution was evaluated by the Cophenetic Correlation Coefficient (CCC), which measures how faithfully the dendrogram preserves pairwise distances; CCC > 0.80 is generally considered a good fit [134]. Accession clusters were defined at k = 3 using the fcluster function with the "maxclust" criterion. Cluster stability was assessed using the mean silhouette score [136], which measures how similar each accession is to its own cluster compared to the nearest neighbouring cluster (range −1 to +1; values > 0.25 indicate meaningful structure). Heatmaps were generated in Python using matplotlib, with dendrograms aligned to heatmap rows and columns. Genes were colored above the heatmap according to their functional group assignment; the focal species *M. dubia* was highlighted with a bold red label. All index values were used without standardization to preserve their native interpretable scales.

Codon Coding and Decoding Mechanisms

The complete tRNA complement was compiled from GenBank annotations across all 23 accessions. A consensus set of 30 unique tRNA loci was established, representing those present in the 23 accessions. Codon decoding categories were assigned based on the mechanistic interaction between each mRNA codon and its cognate chloroplast tRNA anticodon. Category I (44 codons, ~77%) encompasses standard Watson–Crick pairing and G_{34}–U wobble, the canonical framework established by Crick [68] and structurally validated by Agris [69]. Category II (12 codons, ~16%) covers superwobble decoding, in which a single unmodified U_{34}-containing tRNA reads all four synonymous codons in a four-fold degenerate box without an isoacceptor partner; this mechanism was experimentally demonstrated in *Nicotiana tabacum* transplastomic knockouts by Rogalski et al. [61] and Alkatib et al. [70].

The remaining three categories involve post-transcriptional tRNA modifications. Category III (CGU/CGC/CGA, ~3.8%) is decoded by inosine (I_{34}) generated via $A_{34}\rightarrow I_{34}$ deamination of trnR-ACG by the chloroplast-targeted TadA/TADA enzyme [71,72]. Category IV (AUA, ~2.8%) requires $C_{34}\rightarrow k^2C_{34}$ or ava^2C_{34} modification of trnI-CAU, simultaneously converting codon and amino acid specificity from AUG/Met to AUA/Ile [73–75]. Category V (CGG, ~0.3%) is decoded via a two-out-of-three mechanism by unmodified cp-tRNAArg(ACG), bypassing the wobble position entirely [143].

Structural Analysis and Comparative Genomics

Twenty-three Myrtales complete cp genomes (156,129–160,326 bp) were retrieved from NCBI GenBank, spanning four families, and parsed using Biopython v1.79 (Cock et al., 2009). GC content analysis was performed using Geneious Prime® 2025.2.1 (Biomatters Ltd., New Zealand) across the entire genome and in specific regions. Simple sequence repeats (SSRs) were identified across all 23 cp genomes using the SSRs subcommand of CPStools v3.0 [130], with minimum repeat thresholds of 10 units for mononucleotide, 5 for dinucleotide, 4 for trinucleotide, and 3 for tetra-, penta-, and hexanucleotide motifs. Long dispersed repeats, including forward, palindromic (inverted), reverse, and complementary types, were detected using the LSRs subcommand of CPStools v3.0, with a minimum repeat length of 30 bp and a Hamming distance of 3 (allowing up to 3 mismatches).

Nucleotide diversity (π) was calculated across all 23 accessions using the Pi subcommand of CPStools v3.0, applying a sliding window approach to a multiple sequence alignment of the 23 cp genomes constructed with MAFFT v7.490 (FFT-NS-2 algorithm). Three window sizes and step configurations were employed, 100 bp/10 bp, 200 bp/20 bp, and 500 bp/50 bp, to capture diversity patterns at fine, intermediate, and broad genomic scales, respectively. To avoid artificial inflation of π estimates at region boundaries, sliding windows physically overlapping the IRB–SSC and SSC–IRA junctions by more than 500 bp were excluded from the regional analyses, resulting in a working dataset of 3,146 windows derived from the 500 bp/50 bp configuration used for all statistical comparisons

Differences in π among genomic regions (LSC, SSC, IRA, IRB) and functional categories (PCGs, intergenic regions, rRNA genes, tRNA genes) were assessed using the Kruskal-Wallis H test, a non-parametric equivalent of one-way ANOVA appropriate for data that does not conform to normality assumptions, a condition confirmed by the strongly right-skewed distributions observed for IR regions and rRNA genes. Pairwise post-hoc comparisons were conducted using Dunn's test with Bonferroni

correction for multiple comparisons to control the family-wise error rate. Results of pairwise comparisons were summarized using Compact Letter Display (CLD), where groups sharing the same letter are not significantly different at α = 0.05. All statistical analyses were performed in Python using the scipy.stats and scikit-posthocs libraries.

Sliding window analysis of nucleotide sequence variability (SWAN) [144] of 78 PCGs, obtained of the 23 cp genomes, was conducted using a comprehensive framework to examine six key evolutionary metrics: entropy (sequence variability), pairwise variability (pvar), nucleotide diversity (π), non-synonymous substitution rates (Ka), synonymous substitution rates (Ks), and Ka/Ks ratios (selective pressure indicators). Analysis was performed using a Python pipeline with 3-nucleotide windows with 1-nucleotide step intervals across multiple sequence alignments constructed with MAFFT v7.490 (FFT-NS-2 algorithm, --auto flag). The entropy values were calculated using Shannon's information theory formula: $H = -\Sigma(pi \times \log2(pi))$, where pi represents the frequency of each nucleotide at position i. Nucleotide diversity (π) was calculated as the average number of nucleotide differences per site between sequences. Ka and Ks values were estimated using the Nei-Gojobori method with the Jukes-Cantor correction for multiple substitutions.

For statistical analysis of sliding window evolutionary metrics (entropy, pairwise variability, nucleotide diversity π, Ka, Ks, and Ka/Ks ratios), a PCA was performed on the standardized (z-scored) six-metric per-taxon mean matrix (23 × 6). K-means clustering (k = 2–8; n = 50 random initializations) was applied to the 3-PC embedding; optimal k was chosen by maximizing the silhouette coefficient and minimizing the Davies-Bouldin (DB) index. The full 468-feature matrix (23 taxa × 78 genes × 6 metrics, z-scored per feature) was used for k-means and HC. Hierarchical clustering (Ward linkage) was performed separately on each of the six per-metric matrices and on the combined 468-feature matrix, producing seven HC dendrogram heatmaps. Cophenetic correlation coefficients (CPCC) were computed for both the species (row) and gene (column) dendrograms of each heatmap using SciPy. Pairwise genomic distances (Euclidean on the standardized six-metric mean profile) were visualized as a clustermap/distance matrix.

The boundaries of the four structural regions of the cp genome, the large single-copy (LSC) region, inverted repeat B (IRB), small single-copy (SSC) region, and inverted repeat A (IRA), and the four inter-region junction sites, JLB (LSC/IRb), JSB (IRb/SSC), JSA (SSC/IRa), and JLA (IRa/LSC), were analyzed and visualized comparatively across 23 Myrtales cp genomes using CPJSdraw v0.0.1. [145]. Individual GenBank files (.gb format) were used as input. The quadripartite structure of each genome was defined by specifying the start and end coordinates of the LSC, IRb, SSC, and IRa regions in a CPJSdraw configuration file. CPJSdraw was executed using this configuration file as input, and the resulting SVG figure was post-processed to apply Arial font uniformly, add abbreviated species names with GenBank accession numbers, and ensure consistent gene synteny representation across all accessions.

A Circos-based circular visualization of cp genome structure and synteny was use to compare the 23 cp genomes. Seven Circos-formatted input files were generated via custom Python scripts. The karyotype file assigned each plastome as an independent ideogram, color-coded by taxonomic affiliation at the family and tribe level. Gene position tile tracks were extracted directly from GenBank feature tables for CDS, tRNA, and rRNA features. A quantitative GC content track was computed using 5,000 bp sliding windows with a 1,000 bp step, and a gene density histogram was calculated over non-overlapping 10,000 bp bins. A total of 734 CDS collinearity links were generated by matching shared gene loci between *Myrciaria dubia* and ten Myrteae species. Circos v0.69-8 [146] was configured with five concentric data tracks positioned between radii 0.62r and 0.95r, tick marks at 10 and 50 kb intervals, and link rendering rules differentiating clade affiliation by color and z-order. All outputs were rendered at 4,000-pixel resolution and composited into a publication-quality two-panel figure at 600 DPI using Pillow v10.

Pairwise sequence identity of 22 Myrtales cp genomes was visualized against *M. dubia* (PX251231.1) as the reference using mVISTA with the NUCmer alignment algorithm. Sequences were aligned in a sliding window of 200 bp at a minimum identity threshold of 50%. The annotated *M. dubia* cp genome was used to define genomic regions, LSC, IRA, IRB, and SSC, and to classify features as PCGs, tRNA/rRNA, or intergenic regions. All 23 cp genomes, were included to provide a comprehensive order-level comparison.

Phylogenetic analysis

For phylogenetic reconstruction, complete cp genome sequences of 38 species representing major angiosperm orders were obtained from the Chloroplast Genome Information Resource (https://ngdc.cncb.ac.cn/cgir/). The dataset included 23 Myrtales species, 14 species from six additional angiosperm orders, and *Medicago truncatula* as outgroup (Supplementary Table 7). Based on these complete cp genomes, a dataset was constructed, analyzed, and stored in Geneious Prime® 2025.2.1

(Biomatters Ltd., New Zealand). The 38 complete cp genome sequences in FASTA format were uploaded to the web server NGPhylogeny.fr (https://ngphylogeny.fr/). On this online platform, using the workflow maker, the sequences were first aligned using a fast Fourier transform approach (Software MAFFT v7.407_1) [147] with default parameters, and then the aligned sequences were curated to select phylogenetically informative regions using block mapping and gathering with entropy approaches (Software BMGE v1.12_1) [148].

Curated multiple sequence alignments (Supplementary Sequence 1) were input into the IQ-TREE web server (http://iqtree.cibiv.univie.ac.at/) [149]. In this online platform, the best-fit substitution model of evolution for cp genomes was selected using ModelFinder [150], based on three selection criteria: Akaike Information Criterion (AIC), Corrected AIC, and Bayesian Information Criterion (BIC). The three selection criteria identify GTR+F+R4 as the best-fit model. Furthermore, based on the selected model, the evolutionary parameters were optimized, and the maximum likelihood (ML) distances were computed based on the estimated model parameters. One thousand samples were generated for ultrafast bootstrap analysis (random seed number: 355502) for the construction of the ML phylogenetic tree. The consensus phylogenetic tree obtained in Newick format was exported for visualization and editing in iTOL v7 (https://itol.embl.de).

Statistical Analysis

All basic and advanced statistical analyses were performed using Python (version 3.11.6) in the Julius computational platform (https://julius.ai/), which provided cloud-based processing power essential for computationally intensive analyses.

Data availability

The complete cp genome sequence data supporting the findings of this study are publicly available at the NCBI GenBank under accession number PX251231.1 (https://www.ncbi.nlm.nih.gov/nuccore/PX251231.1). The associated Bioproject, BioSample, and SRA numbers are PRJNA615000, SAMN50748728, and SRR35115586, respectively. All comparative cp genome sequences used in this study were obtained from NCBI GenBank and are comprehensively listed in Supplementary Table 7 with their accession numbers, genome sizes, and direct NCBI URLs for accessibility. All other data generated or analyzed during this study are included in this published article and its supplementary information files.

References

1. Daniell, H., Lin, C.-S., Yu, M. & Chang, W.-J. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biol* **17**, 134 (2016).
2. Wicke, S., Schneeweiss, G. M., dePamphilis, C. W., Müller, K. F. & Quandt, D. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Mol Biol* **76**, 273–297 (2011).
3. Dobrogojski, J., Adamiec, M. & Luciński, R. The chloroplast genome: a review. *Acta Physiol Plant.* **42**, 1–13 (2020).
4. Tonti-Filippini, J., Nevill, P. G., Dixon, K. & Small, I. What can we do with 1000 plastid genomes? *Plant J.* **90**, 808–818 (2017).
5. Gitzendanner, M. A., Soltis, P. S., Wong, G. K.-S., Ruhfel, B. R. & Soltis, D. E. Plastid phylogenomic analysis of green plants: A billion years of evolutionary history. *Am. J. Bot.* **105**, 291–301 (2018).
6. Hua, Z. *et al.* Towards comprehensive integration and curation of chloroplast genomes. *Plant Biotechnol J* **20**, 2239–2241 (2022).
7. Jansen, R. K. *et al.* Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *PNAS* **104**, 19369–19374 (2007).
8. Asif, H. *et al.* The chloroplast genome sequence of *Syzygium cumini* (L.) and its relationship with other angiosperms. *Tree Genetics & Genomes* **9**, 867–877 (2013).
9. Shaw, J., Lickey, E. B., Schilling, E. E. & Small, R. L. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. *American Journal of Botany* **94**, 275–288 (2007).
10. Dong, W. *et al.* ycf1, the most promising plastid DNA barcode of land plants. *Sci Rep* **5**, 8348 (2015).
11. Biffin, E., Craven, L. A., Crisp, M. D. & Gadek, P. A. Molecular systematics of Syzygium and allied genera (Myrtaceae): evidence from the chloroplast genome. *Taxon* **55**, 79–94 (2006).

12. Wilson, P. G., O'Brien, M. M., Gadek, P. A. & Quinn, C. J. Myrtaceae revisited: a reassessment of infrafamilial groups. *Am. J. Bot.* **88**, 2013–2025 (2001).
13. Biffin, E. *et al.* Evolution of exceptional species richness among lineages of fleshy-fruited Myrtaceae. *Ann Bot* **106**, 79–93 (2010).
14. Lucas, E. J. *et al.* Suprageneric phylogenetics of Myrteae, the generically richest tribe in Myrtaceae (Myrtales). *Taxon* **56**, 1105–1128 (2007).
15. Vasconcelos, T. N. C. *et al.* Myrteae phylogeny, calibration, biogeography and diversification patterns: Increased understanding in the most species rich tribe of Myrtaceae. *Mol Phylogenet Evol* **109**, 113–137 (2017).
16. Hughes, C. E., Pennington, R. T. & Antonelli, A. Neotropical Plant Evolution: Assembling the Big Picture. *Bot J Linn Soc* **171**, 1–18 (2013).
17. Akter, M. S., Oh, S., Eun, J. B. & Ahmed, M. Nutritional compositions and health promoting phytochemicals of camu-camu (*Myrciaria dubia*) fruit: A review. *Food Research International* **44**, 1728–1732 (2011).
18. Imán Correa, S., Pinedo Freyre, S. & Melchor Aldana, M. Morphologic characterization and evaluation of the national collection of germoplasm of camu camu *Myrciaria dubia* (H.B.K) Mc Vaugh, of the INIA Loreto-Peru. *Sci. agropecu* 189–201 (2011) doi:10.17268/sci.agropecu.2011.04.01.
19. Castro, J., Maddox, J., Cobos, M. & Imán, S. *Myrciaria dubia* "Camu Camu" Fruit: Health-Promoting Phytochemicals and Functional Genomic Characteristics. in *Breeding and Health Benefits of Fruit and Nut Crops* 158 (IntechOpen, London, England, 2018).
20. Chirinos, R., Galarza, J., Betalleluz-Pallardel, I., Pedreschi, R. & Campos, D. Antioxidant compounds and antioxidant capacity of Peruvian camu camu (*Myrciaria dubia* (H.B.K.) McVaugh) fruit at different maturity stages. *Food Chemistry* **120**, 1019–1024 (2010).
21. Langley, P. C., Pergolizzi, J. V., Taylor, R. & Ridgway, C. Antioxidant and Associated Capacities of Camu Camu (*Myrciaria dubia*): A Systematic Review. *J Altern Complement Med* **21**, 8–14 (2015).
22. Azevedo, L. *et al.* Camu-camu (Myrciaria dubia) from commercial cultivation has higher levels of bioactive compounds than native cultivation (Amazon Forest) and presents antimutagenic effects in vivo. *Journal of the Science of Food and Agriculture* **99**, 624–631 (2019).
23. Fracassetti, D., Costa, C., Moulay, L. & Tomás-Barberán, F. A. Ellagic acid derivatives, ellagitannins, proanthocyanidins and other phenolics, vitamin C and antioxidant capacity of two powder products from camu-camu fruit (*Myrciaria dubia*). *Food Chemistry* **139**, 578–588 (2013).
24. Neri-Numa, I. A., Soriano Sancho, R. A., Pereira, A. P. A. & Pastore, G. M. Small Brazilian wild fruits: Nutrients, bioactive compounds, health-promotion properties and commercial interest. *Food Research International* **103**, 345–360 (2018).
25. Aguiar, J. P. L. & Souza, F. das C. do A. Camu-Camu (*Myrciaria dubia* HBK): Yogurt Processing, Formulation, and Sensory Assessment. *American Journal of Analytical Chemistry* **06**, 377–381 (2015).
26. Castro, J. C. *et al.* Development and Application of Microsatellite Markers for Genetic Diversity Assessment and Construction of a Core Collection of *Myrciaria dubia* (Kunth) McVaugh Germplasm from the Peruvian Amazon. *Forests* **15**, 22–22 (2024).
27. Šmíd, J. *et al.* Morphological and genetic diversity of camu-camu [*Myrciaria dubia* (Kunth) McVaugh] in the Peruvian Amazon. *PLOS ONE* **12**, e0179886 (2017).
28. Coulibaly, D. *et al.* Comparative Analysis of Complete Chloroplast Genome and Phenotypic Characteristics of Japanese Apricot Accessions. *Horticulturae* **8**, 794 (2022).
29. Palmer, J. D. Chloroplast DNA exists in two orientations. *Nature* **301**, 92–93 (1983).
30. Zhu, A., Guo, W., Gupta, S., Fan, W. & Mower, J. P. Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. *New Phytologist* **209**, 1747–1756 (2016).

31. Weng, M.-L., Blazier, J. C., Govindu, M. & Jansen, R. K. Reconstruction of the Ancestral Plastid Genome in Geraniaceae Reveals a Correlation between Genome Rearrangements, Repeats, and Nucleotide Substitution Rates. *Mol Biol Evol* **31**, 645–659 (2014).
32. Khakhlova, O. & Bock, R. Elimination of deleterious mutations in plastid genomes by gene conversion. *The Plant Journal* **46**, 85–94 (2006).
33. Birky, C. W. Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. *Proceedings of the National Academy of Sciences* **92**, 11331–11338 (1995).
34. Howe, C. J. *et al.* Evolution of the chloroplast genome. *Philos Trans R Soc Lond B Biol Sci* **358**, 99–107 (2003).
35. Cauz-Santos, L. A. Beyond conservation: the landscape of chloroplast genome rearrangements in angiosperms. *New Phytologist* **247**, 2571–2580 (2025).
36. Singh, R., Ming, R. & Yu, Q. Comparative Analysis of GC Content Variations in Plant Genomes. *Tropical Plant Biol.* **9**, 136–149 (2016).
37. Shimda, H. & Sugiuro, M. Fine structural features of the chloroplast genome: comparison of the sequenced chloroplast genomes. *Nucleic Acids Res* **19**, 983–995 (1991).
38. Knight, R. D., Freeland, S. J. & Landweber, L. F. Selection, history and chemistry: the three faces of the genetic code. *Trends in Biochemical Sciences* **24**, 241–247 (1999).
39. Freeland, S. J., Knight, R. D., Landweber, L. F. & Hurst, L. D. Early Fixation of an Optimal Genetic Code. *Mol Biol Evol* **17**, 511–518 (2000).
40. Clegg, M. T., Gaut, B. S., Learn, G. H. & Morton, B. R. Rates and patterns of chloroplast DNA evolution. *Proceedings of the National Academy of Sciences* **91**, 6795–6801 (1994).
41. Zhang, R. & Zhang, C.-T. Isochore Structures in the Genome of the Plant *Arabidopsis thaliana*. *J Mol Evol* **59**, 227–238 (2004).
42. Šmarda, P. & Bureš, P. The Variation of Base Composition in Plant Genomes. in *Plant Genome Diversity Volume 1: Plant Genomes, their Residents, and their Evolutionary Dynamics* (eds Wendel, J. F., Greilhuber, J., Dolezel, J. & Leitch, I. J.) 209–235 (Springer, Vienna, 2012). doi:10.1007/978-3-7091-1130-7_14.
43. Costantini, M. & Musto, H. The Isochores as a Fundamental Level of Genome Structure and Organization: A General Overview. *J Mol Evol* **84**, 93–103 (2017).
44. Morton, B. R. The Role of Context-Dependent Mutations in Generating Compositional and Codon Usage Bias in Grass Chloroplast DNA. *J Mol Evol* **56**, 616–629 (2003).
45. Gantt, J. S., Baldauf, S. L., Calie, P. J., Weeden, N. F. & Palmer, J. D. Transfer of rpl22 to the nucleus greatly preceded its loss from the chloroplast and involved the gain of an intron. *EMBO J* **10**, 3073–3078 (1991).
46. Millen, R. S. *et al.* Many Parallel Losses of infA from Chloroplast DNA during Angiosperm Evolution with Multiple Independent Transfers to the Nucleus. *Plant Cell.* **13**, 645–658 (2001).
47. Martin, W. *et al.* Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proceedings of the National Academy of Sciences* **99**, 12246–12251 (2002).
48. Timmis, J. N., Ayliffe, M. A., Huang, C. Y. & Martin, W. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet* **5**, 123–135 (2004).
49. Archibald, J. M. Genomic perspectives on the birth and spread of plastids. *Proceedings of the National Academy of Sciences* **112**, 10147–10153 (2015).
50. Kleine, T., Maier, U. G. & Leister, D. DNA Transfer from Organelles to the Nucleus: The Idiosyncratic Genetics of Endosymbiosis. *Annual Review of Plant Biology* **60**, 115–138 (2009).
51. Rousseau-Gueutin, M. *et al.* Potential Functional Replacement of the Plastidic Acetyl-CoA Carboxylase Subunit (accD) Gene by Recent Transfers to the Nucleus in Some Angiosperm Lineages. *Plant Physiol* **161**, 1918–1929 (2013).
52. Smith, D. R. & Keeling, P. J. Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. *Proceedings of the National Academy of Sciences* **112**, 10177–10184 (2015).

53. Morton, B. R. Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages. *J Mol Evol* **46**, 449–459 (1998).
54. Morton, B. R. & Clegg, M. T. Neighboring base composition is strongly correlated with base substitution bias in a region of the chloroplast genome. *J Mol Evol* **41**, 597–603 (1995).
55. Bulmer, M. The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129**, 897–907 (1991).
56. Ikemura, T. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* **2**, 13–34 (1985).
57. Morton, B. R. Context-Dependent Mutation Dynamics, Not Selection, Explains the Codon Usage Bias of Most Angiosperm Chloroplast Genes. *J Mol Evol* **90**, 17–29 (2022).
58. Parvathy, S. T., Udayasuriyan, V. & Bhadana, V. Codon usage bias. *Mol Biol Rep* **49**, 539–565 (2022).
59. Morton, B. R. Chloroplast DNA codon use: Evidence for selection at the psb A locus based on tRNA availability. *J Mol Evol* **37**, 273–280 (1993).
60. Sugiura, M., Hirose, T. & Sugita, M. Evolution and Mechanism of Translation in Chloroplasts. *Annual Review of Genetics* **32**, 437–459 (1998).
61. Rogalski, M., Karcher, D. & Bock, R. Superwobbling facilitates translation with reduced tRNA sets. *Nat Struct Mol Biol* **15**, 192–198 (2008).
62. Zoschke, R. & Bock, R. Chloroplast Translation: Structural and Functional Organization, Operational Control, and Regulation. *Plant Cell* **30**, 745–770 (2018).
63. Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet* **12**, 32–42 (2011).
64. Hershberg, R. & Petrov, D. A. Selection on codon bias. *Annu Rev Genet* **42**, 287–299 (2008).
65. Hao, J. *et al.* Chloroplast gene expression level is negatively correlated with evolutionary rates and selective pressure while positively with codon usage bias in *Ophioglossum vulgatum* L. *BMC Plant Biol* **22**, 580 (2022).
66. Liu, Q. & Xue, Q. Comparative studies on codon usage pattern of chloroplasts and their host nuclear genes in four plant species. *J Genet* **84**, 55–62 (2005).
67. Nie, X. *et al.* Comparative analysis of codon usage patterns in chloroplast genomes of the Asteraceae family. *Plant Mol Biol Rep* **32**, 828–840 (2014).
68. Crick, F. H. Codon-anticodon pairing: the wobble hypothesis. *J Mol Biol* **19**, 548–555 (1966).
69. Agris, P. F. Wobble position modified nucleosides evolved to select transfer RNA codon recognition: A modified-wobble hypothesis. *Biochimie* **73**, 1345–1349 (1991).
70. Alkatib, S. *et al.* The Contributions of Wobbling and Superwobbling to the Reading of the Genetic Code. *PLoS Genet.* **8**, e1003076 (2012).
71. Karcher, D. & Bock, R. Identification of the chloroplast adenosine-to-inosine tRNA editing enzyme. *RNA* **15**, 1251–1257 (2009).
72. Delannoy, E. *et al.* Arabidopsis tRNA Adenosine Deaminase Arginine Edits the Wobble Nucleotide of Chloroplast tRNAArg(ACG) and Is Essential for Efficient Chloroplast Translation. *Plant Cell* **21**, 2058–2071 (2009).
73. Miyauchi, K. *et al.* A tRNA modification with aminovaleramide facilitates AUA decoding in protein synthesis. *Nat Chem Biol* **21**, 522–531 (2025).
74. Muramatsu, T. *et al.* A novel lysine-substituted nucleoside in the first position of the anticodon of minor isoleucine tRNA from Escherichia coli. *Journal of Biological Chemistry* **263**, 9261–9267 (1988).
75. Muramatsu, T. *et al.* Codon and amino-acid specificities of a transfer RNA are both converted by a single post-transcriptional modification. *Nature* **336**, 179–181 (1988).
76. Wolfe, K. H., Li, W. H. & Sharp, P. M. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences* **84**, 9054–9058 (1987).
77. Ravi, V., Khurana, J. P., Tyagi, A. K. & Khurana, P. An update on chloroplast genomes. *Plant Syst Evol* **271**, 101–122 (2008).

78. Perry, A. S. & Wolfe, K. H. Nucleotide Substitution Rates in Legume Chloroplast DNA Depend on the Presence of the Inverted Repeat. *J Mol Evol* **55**, 501–508 (2002).
79. Wicke, S., Schneeweiss, G. M., dePamphilis, C. W., Müller, K. F. & Quandt, D. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Mol Biol* **76**, 273–297 (2011).
80. Petit, R. J. *et al.* Comparative organization of chloroplast, mitochondrial and nuclear diversity in plant populations. *Mol Ecol* **14**, 689–701 (2005).
81. Shaw, J., Lickey, E. B., Schilling, E. E. & Small, R. L. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. *American Journal of Botany* **94**, 275–288 (2007).
82. Avise, J. C. Phylogeography: retrospect and prospect. *Journal of Biogeography* **36**, 3–15 (2009).
83. de Andrade Wagner, M., Bogoni, J. A. & Fiaschi, P. Myrtaceae richness and distribution across the Atlantic Forest Domain are constrained by geoclimatic variables. *Plant Ecol* **223**, 1079–1092 (2022).
84. De Souza Neto, J. D. *et al.* Advances and perspectives on the evolutionary history and diversification of Neotropical Myrteae (Myrtaceae). *Bot J Linn Soc* **199**, 173–195 (2022).
85. Comes, H. P. & Kadereit, J. W. The effect of Quaternary climatic changes on plant distribution and evolution. *Trends in Plant Science* **3**, 432–438 (1998).
86. Jackson, S. T. & Overpeck, J. T. Responses of plant populations and communities to environmental changes of the late Quaternary. *Paleobiology* **26**, 194–220 (2000).
87. Giaretta, A., de Menezes, L. F. T. & Peixoto, A. L. Diversity of Myrtaceae in the southeastern Atlantic forest of Brazil as a tool for conservation. *Braz. J. Bot* **38**, 175–185 (2015).
88. Yan, M.-H. *et al.* Phylogeography of the Endangered Endemic *Perkinsiodendron macgregorii* Based on Chloroplast Genome Data. *Diversity* **17**, 439 (2025).
89. Li, L.-L. *et al.* The Ka /Ks and πa /πs Ratios under Different Models of Gametophytic and Sporophytic Selection. *Genome Biol Evol* **15**, evad151 (2023).
90. Qi, C. *et al.* Fixation of Expression Divergences by Natural Selection in Arabidopsis Coding Genes. *Int J Mol Sci* **25**, 13710 (2024).
91. Allen, J. F. *et al.* Evolution of the chloroplast genome. *Philos Trans R Soc Lond B Biol Sci* **358**, 99–107 (2003).
92. Drescher, A., Ruf, S., Calsa, T., Carrer, H. & Bock, R. The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. *Plant J* **22**, 97–104 (2000).
93. Vogel, J., Hübschmann, T., Börner, T. & Hess, W. R. Splicing and intron-internal RNA editing of *trnK-matK* transcripts in barley plastids: support for MatK as an essential splice factor1. *Journal of Molecular Biology* **270**, 179–187 (1997).
94. Martín, M. & Sabater, B. Plastid *ndh* genes in plant evolution. *Plant Physiology and Biochemistry* **48**, 636–645 (2010).
95. Li, Y.-C., Korol, A. B., Fahima, T., Beiles, A. & Nevo, E. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Molecular Ecology* **11**, 2453–2465 (2002).
96. Provan, J., Powell, W. & Hollingsworth, P. M. Chloroplast microsatellites: new tools for studies in plant ecology and evolution. *Trends in Ecology & Evolution* **16**, 142–147 (2001).
97. Schlötterer, C. & Tautz, D. Slippage synthesis of simple sequence DNA. *Nucleic Acids Res* **20**, 211–215 (1992).
98. Vinces, M. D., Legendre, M., Caldara, M., Hagihara, M. & Verstrepen, K. J. Unstable tandem repeats in promoters confer transcriptional evolvability. *Science* **324**, 1213–1216 (2009).
99. Ogihara, Y., Terachi, T. & Sasakuma, T. Intramolecular recombination of chloroplast genome mediated by short direct-repeat sequences in wheat species. *Proceedings of the National Academy of Sciences* **85**, 8573–8577 (1988).

100. Haberle, R. C., Fourcade, H. M., Boore, J. L. & Jansen, R. K. Extensive Rearrangements in the Chloroplast Genome of *Trachelium caeruleum* Are Associated with Repeats and tRNA Genes. *J Mol Evol* **66**, 350–361 (2008).
101. Fotsing, S. F. *et al.* The impact of short tandem repeat variation on gene expression. *Nat Genet* **51**, 1652–1659 (2019).
102. George, B., Bhatt, B. S., Awasthi, M., George, B. & Singh, A. K. Comparative analysis of microsatellites in chloroplast genomes of lower and higher plants. *Curr Genet* **61**, 665–677 (2015).
103. Verma, D. & Daniell, H. Chloroplast Vector Systems for Biotechnology Applications. *Plant Physiology* **145**, 1129–1143 (2007).
104. Carvalho, L. R. *et al.* The complete chloroplast genome sequence of *Eugenia klotzschiana* O. Berg unveils the evolutionary dynamics in plastomes of Myrteae DC. Tribe (Myrtaceae). *Gene* **876**, 147488 (2023).
105. Bayly, M. J. *et al.* Chloroplast genome analysis of Australian eucalypts – *Eucalyptus*, *Corymbia*, *Angophora*, *Allosyncarpia* and *Stockwellia* (Myrtaceae). *Molecular Phylogenetics and Evolution* **69**, 704–716 (2013).
106. Wilson, P. G., O'Brien, M. M., Gadek, P. A. & Quinn, C. J. Myrtaceae revisited: a reassessment of infrafamilial groups. *Am J Bot* **88**, 2013–2025 (2001).
107. Lucas, E. J. *et al.* Suprageneric phylogenetics of Myrteae, the generically richest tribe in Myrtaceae (Myrtales). *TAXON* **56**, 1105–1128 (2007).
108. Landrum, L. R. & Kawasaki, M. L. The genera of Myrtaceae in Brazil: an illustrated synoptic treatment and identification keys. *Brittonia* **49**, 508–536 (1997).
109. Santos, M. F., Sano, P. T., Forest, F. & Lucas, E. Phylogeny, morphology and circumscription of Myrcia sect. Sympodiomyrcia (Myrcia s.l., Myrtaceae). *TAXON* **65**, 759–774 (2016).
110. Thornhill, A. H., Ho, S. Y. W., Külheim, C. & Crisp, M. D. Interpreting the modern distribution of Myrtaceae using a dated molecular phylogeny. *Mol Phylogenet Evol* **93**, 29–43 (2015).
111. Biffin, E. *et al.* Evolution of exceptional species richness among lineages of fleshy-fruited Myrtaceae. *Ann Bot* **106**, 79–93 (2010).
112. Grattapaglia, D. *et al.* Progress in Myrtaceae genetics and genomics: Eucalyptus as the pivotal genus. *Tree Genetics & Genomes* **8**, 463–508 (2012).
113. Thornhill, A. H. & Macphail, M. Fossil myrtaceous pollen as evidence for the evolutionary history of Myrtaceae: A review of fossil *Myrtaceidites* species. *Review of Palaeobotany and Palynology* **176–177**, 1–23 (2012).
114. Antonelli, A. *et al.* Amazonia is the primary source of Neotropical biodiversity. *Proc Natl Acad Sci U S A* **115**, 6034–6039 (2018).
115. Hoorn, C. *et al.* Amazonia through time: Andean uplift, climate change, landscape evolution, and biodiversity. *Science* **330**, 927–931 (2010).
116. Rull, V. Neotropical biodiversity: timing and potential drivers. *Trends Ecol Evol* **26**, 508–513 (2011).
117. Leite, R. N. & Rogers, D. S. Revisiting Amazonian phylogeography: insights into diversification hypotheses and novel perspectives. *Org Divers Evol* **13**, 639–664 (2013).
118. Winter, M., Devictor, V. & Schweiger, O. Phylogenetic diversity and nature conservation: where are we? *Trends Ecol Evol* **28**, 199–204 (2013).
119. Forest, F. *et al.* Preserving the evolutionary potential of floras in biodiversity hotspots. *Nature* **445**, 757–760 (2007).
120. Balbinott, N., Rodrigues, N. F., Guzman, F. L., Turchetto-Zolet, A. C. & Margis, R. Perspectives in Myrtaceae evolution from plastomes and nuclear phylogenies. *Genet Mol Biol* **45**, e20210191 (2022).
121. Vasconcelos, T. N. C. *et al.* Myrteae phylogeny, calibration, biogeography and diversification patterns: Increased understanding in the most species rich tribe of Myrtaceae. *Mol Phylogenet Evol* **109**, 113–137 (2017).

122. Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phyt. Bull.* **19**, 11–15 (1987).
123. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
124. Jalili, V. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Research* **48**, W395–W402 (2020).
125. Dierckxsens, N., Mardulyn, P. & Smits, G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* **45**, e18 (2017).
126. Jin, J.-J. *et al.* GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol* **21**, 241 (2020).
127. Tillich, M. *et al.* GeSeq - versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* **45**, W6–W11 (2017).
128. Chan, P. P. & Lowe, T. M. tRNAscan-SE: Searching for tRNA genes in genomic sequences. *Methods Mol Biol* **1962**, 1–14 (2019).
129. Greiner, S., Lehwark, P. & Bock, R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res* **47**, W59–W64 (2019).
130. Huang, L., Yu, H., Wang, Z. & Xu, W. CPStools: A package for analyzing chloroplast genome sequences. *IMetaOmics* **1**, e25 (2024).
131. Sharp, P. M. & Li, W. H. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**, 1281–1295 (1987).
132. Wright, F. The 'effective number of codons' used in a gene. *Gene* **87**, 23–29 (1990).
133. Härdle, W. K., Simar, L. & Fengler, M. R. Cluster Analysis. in *Applied Multivariate Statistical Analysis* (eds Härdle, W. K., Simar, L. & Fengler, M. R.) 373–406 (Springer International Publishing, Cham, 2024). doi:10.1007/978-3-031-63833-6_13.
134. Sokal, R. R. & Rohlf, F. J. *Biometry: The Principles and Practices of Statistics in Biological Research*. (W. H. Freeman, New York, 1995).
135. Legendre, P. & Legendre, L. *Numerical Ecology*. (Elsevier, Amsterdam, 2012).
136. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 53–65 (1987).
137. Sun, X., Yang, Q. & Xia, X. An Improved Implementation of Effective Number of Codons (Nc). *Mol Biol Evol* **30**, 191–196 (2013).
138. Novembre, J. A. Accounting for background nucleotide composition when measuring codon usage bias. *Mol Biol Evol* **19**, 1390–1394 (2002).
139. Sueoka, N. Directional mutation pressure and neutral molecular evolution. *Proc Natl Acad Sci U S A* **85**, 2653–2657 (1988).
140. Moriyama, E. N. & Powell, J. R. Codon usage bias and tRNA abundance in Drosophila. *J Mol Evol* **45**, 514–523 (1997).
141. Huo, X. *et al.* Analysis of synonymous codon usage of transcriptome database in Rheum palmatum. *PeerJ* **9**, e10450 (2021).
142. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**, 14863–14868 (1998).
143. Delannoy, E. *et al.* Arabidopsis tRNA Adenosine Deaminase Arginine Edits the Wobble Nucleotide of Chloroplast tRNAArg(ACG) and Is Essential for Efficient Chloroplast Translation. *Plant Cell* **21**, 2058–2071 (2009).
144. Proutski, V. & Holmes, E. SWAN: sliding window analysis of nucleotide sequence variability. *Bioinformatics* **14**, 467–468 (1998).
145. Li, H. *et al.* CPJSdraw: analysis and visualization of junction sites of chloroplast genomes. *PeerJ* **11**, e15326 (2023).
146. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639–1645 (2009).

147. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* **30**, 772–780 (2013).
148. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).
149. Wong, T. K. F. *et al.* IQ-TREE 3: Phylogenomic Inference Software using Complex Evolutionary Models. https://ecoevorxiv.org/repository/view/8916/ (2025).
150. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**, 587–589 (2017).

Acknowledgements
The authors thank Lolo Lumba, Angel Vizcarra, Américo Tuesta, and Celsio Gonzales for their assistance in the field and sample collection. The authors also thank the Field Museum Pritzker Laboratory for Molecular Systematics and Evolution Core Facility for providing access to their equipment for molecular analysis.
Funding
This research was supported by the National Council for Science, Technology, and Technological Innovation (CONCYTEC) through the PROCIENCIA program under the Basic Research Projects funding initiative “E041-2024-03” (Contract Number PE501088786-2024-PROCIENCIA). We also acknowledge the Universidad Nacional de la Amazonía Peruana (UNAP) for institutional support through the approval of research projects under the Rectoral Resolutions N° 0686-2015-UNAP and N° 0449-2024-UNAP.

Author information
Authors and Affiliations

Specialized Unit of Biotechnology Research Laboratory, Natural Resources Research Center of UNAP, Universidad Nacional de la Amazonía Peruana (UNAP), Iquitos 16001, Peru.
Juan C. Castro, Marianela Cobos, Pedro M. Adrianzen, Carlos G. Castro, Hicler N. Rodriguez, Jorge L. Marapara
Academic Department of Biomedical Sciences and Biotechnology, Faculty of Biological Sciences, Universidad Nacional de la Amazonía Peruana (UNAP), Iquitos 16001, Peru
Juan C. Castro, Marianela Cobos, Pedro M. Adrianzen, Hicler N. Rodriguez, Jorge L. Marapara
Institute of Physics of São Carlos, Universidade de São Paulo, Street João Dagnone 1100, São Carlos, SP 13563-723, Brazil
Diego A. Leonardo, Jhon A. Vargas
Academic Department of Ecology and Conservation, Faculty of Forestry Sciences, Universidad Nacional de la Amazonía Peruana (UNAP), Iquitos 16001, Peru
Sixto A. Imán
Direction of Genetic Resources and Biotechnology, San Roque Agricultural Experimental Station, Instituto Nacional de Innovación Agraria (INIA), Iquitos 16001, Peru
Sixto A. Imán
Pritzker Laboratory for Molecular Systematics and Evolution, Field Museum of Natural History, 1400 S. Lake Shore Drive, Chicago, IL 60605, USA
J. Dylan Maddox

Contributions
J.C.C. conceived, acquired funds, and supervised the project, performed data curation, validated, and conducted basis and advanced bioinformatic and biostatistical analysis, wrote the original draft. D.A.L. and J.A.V. conceptualized and conducted the *in-silico* analysis to determine the three-dimensional structures of protein complexes, wrote, reviewed, and edited the manuscript. All authors have read and approved the manuscript. M.C. obtained resources, supervised the project, wrote, reviewed, and edited the manuscript, and validated and performed data curation. P.M.A. performed the design of experiments, and analyses, wrote, reviewed, and edited the manuscript, and performed bioinformatic analysis. C.G.C. performed the experimental design, purified genomic DNA, revised, criticized, and edited the manuscript. J.D.P. performed the construction and sequencing of libraries, wrote, reviewed, and edited the manuscript. S.A.I. collected the botanical samples, performed the experimental design, revised, criticized, and edited the manuscript. J.L.M. conducted plant material identification, revised, criticized, and edited the manuscript. H.N.R. performed the experimental design, purified genomic DNA, revised, criticized, and edited the manuscript. J.D.M. supervised the project, performed data curation, performed basic and advanced bioinformatic and biostatistical analysis, wrote, reviewed, and edited the manuscript.

Corresponding authors

Correspondences to Juan C. Castro (juan.castro@unapiquitos.edu.pe) and J. Dylan Maddox (dmaddox@fieldmuseum.org)

Ethics declarations

Competing interests

The authors declare no competing interests.

Additional information

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Rights and permissions