Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Phylogeny-associated codon evolution and conditional boundary dinucleotide structure in Platycerium plastomes

Jun-Wei Yang

or0706555@gmail.com

Independent Researcher

Cheng-Hsun Lee

Kuan Mei Clinic

Hsuan-Wei Huang

Kuan Mei Clinic

Wei-Mei Hsieh

Xiang Le Zhu Yi

Su-Lan Hsu

Xiang Le Zhu Yi

Research Article

Keywords: codon context, fern evolution, organelle genome, phylogenetic signal, sequence composition, synonymous codon usage

Posted Date: July 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10454969/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Synonymous variation in plastid genes may record both lineage history and local sequence-context constraints. We integrated phylogenetic codon-use analysis with exact-codon conditional randomization across 18 Platycerium species. Eighty-four reference coding features were reconstructed from a published whole-plastome alignment and audited; the primary matrix comprised 44 editing-free, intact genes and 792 species–gene sequences. Relative synonymous codon usage (RSCU) distance increased with plastome-tree distance, and an RSCU phenogram recovered the three recognized geographic lineages, indicating that codon profiles retain phylogenetic structure despite narrow ranges of genomic GC content and effective number of codons. Gene-level RSCU divergence decreased with coding-sequence length, emphasizing the need to distinguish short-gene sampling effects from evolutionary differentiation. For each coding sequence, 9,999 randomizations retained protein sequence, the exact number of every codon, nucleotide composition, and within-codon dinucleotides while changing synonymous-codon placement. Codon-boundary CpG was lower than its conditional expectation, with negative median effects in all species and most genes and persistence under reconstruction and length sensitivities. TpA showed a pooled deficit but greater gene heterogeneity. Analysis of all 16 boundary dinucleotides placed CpG within a broader redistribution of neighbouring codon contexts. Together, the results connect lineage-associated synonymous divergence with a conserved codon-order tendency and support an evolutionary model in which shared plastid composition, ancestry, gene-level constraints, and local sequence context jointly shape codon organization.

Introduction

Plastid genomes are compact, gene-rich chromosomes whose coding sequences retain information about organelle ancestry, lineage diversification, mutation and repair, and constraints acting on transcription and translation (Wicke et al. 2011; Daniell et al. 2016). Synonymous substitutions do not alter amino-acid identity, but they are not evolutionarily featureless. Their accumulation reflects the available mutation spectrum, genetic drift, background nucleotide composition, and possible constraints on translational efficiency or accuracy (Sharp and Li 1986; Wright 1990; Suzuki and Morton 2016). A species-level codon profile may therefore contain a weak but detectable record of shared ancestry even when the plastome is compositionally conservative.

Phylogenetic interpretation is nevertheless challenging. Closely related species inherit similar plastid genomes, so a correlation between codon-use distance and sequence-tree distance is not evidence that codon bias caused lineage divergence. Conversely, failure of a codon-use phenogram to reproduce every sequence-tree branch does not make codon information biologically irrelevant: synonymous composition can retain major lineage structure while also responding to gene-specific histories. Comparing a codon-use phenogram with an independently defined plastome tree, and repeating the association within major lineages, can distinguish a broad phylogenetic signal from separation driven only by the deepest clades (Mantel 1967; Robinson and Foulds 1981).

Gene-level codon metrics introduce additional complications. Relative synonymous codon usage (RSCU) is estimated within synonymous families, so short genes provide few observations and a small number of substitutions can generate a large apparent distance. Effective number of codons (ENC), GC at synonymous third positions, neutrality plots, and parity-rule-2 summaries are often used to describe mutational and selective influences, but no single diagnostic uniquely estimates selection (Wright 1990; Suzuki and Morton 2016). RNA editing further separates genomic codons from mature plastid transcripts. Annotation-aware filtering, gene-length sensitivity, and cautious interpretation are consequently essential when codon evolution is compared across plastomes.

Dinucleotides provide another level of sequence organization. CpG and TpA occur non-randomly in plant genomes and plastids, and context-dependent mutation can contribute to compositional and codon-use patterns in chloroplast DNA (Boudraa and Perrin 1987; Morton 2003, 2023). A raw deficit does not demonstrate that adjacent codons are ordered non-randomly. It can arise because a coding sequence uses few C-ending or G-starting codons, because its protein requires particular amino acids, or because its synonymous codon inventory has a particular nucleotide composition. Related confounding between codon-pair statistics and dinucleotide frequencies has been demonstrated in other coding systems (Tulloch et al. 2014; Kunec and Osterrieder 2016). A conditional null model is therefore required to ask whether synonymous-codon placement still changes the dinucleotide spanning a codon boundary after protein sequence and exact codon use are fixed.

The staghorn fern genus *Platycerium* contains Afro-American, Javan-Australian, and Malayan-Asian lineages supported by plastid phylogenomics and biogeographic analyses (Xue et al. 2024; Ciarle et al. 2025). Public whole-plastome sequences and a published branch-length tree spanning these lineages provide a compact genus-wide system for connecting broad codon evolution with local codon-order structure. The same taxa can be examined at two complementary scales: divergence among whole-species RSCU profiles and conditional sequence context within homologous genes.

We therefore asked four questions. First, does RSCU dissimilarity increase with plastome evolutionary distance and recover the three recognized lineages? Second, how do overall composition, ENC, third-position asymmetry, and coding-sequence length influence gene-level interpretation? Third, are codon-boundary CpG and TpA counts lower than expected after synonymous codons are conditionally reordered? Fourth, are the boundary effects consistent across species, genes, lineages, reconstruction choices, and gene-length thresholds? All 16 boundary dinucleotides were analysed to prevent selective reporting, and permutation draws were not treated as biological replicates. This integrated design distinguishes lineage-associated codon divergence from codon-order signals that remain after protein sequence, exact codon counts, nucleotide composition, and within-codon dinucleotides are held constant.

Materials and methods

Public plastomes and homology-guided coding-sequence reconstruction

We analysed the 18 *Platycerium* plastomes represented in the public alignment accompanying Ciarle et al. (2025). The accessions were OR601544.1–OR601558.1, OR601560.1–OR601561.1, and MN623367.1. After gaps were removed, every alignment row was sequence-identical to its stated GenBank record. The sample comprised seven Afro-American, four Javan-Australian, and seven Malayan-Asian species.

Coding sequences were reconstructed by mapping the 84 unique protein-coding features of the annotated *P. bifurcatum* record MN623367.1 to its sequence-identical alignment row and then transferring the aligned feature intervals to each species. Join and reverse-strand feature parts were reconstructed in feature order, alignment gaps were removed, and reading frame, length modulo three, terminal stop, and in-frame genomic stop triplets were recorded for every species–gene combination. This procedure is a homology-guided reconstruction from a published whole-plastome alignment; it is not direct extraction from independently annotated GenBank features for all 18 taxa.

Of the 1,512 reconstructed species–gene combinations, 1,509 passed the coding-integrity audit. Three *rpl32* reconstructions failed and this gene was excluded. Thirty-nine reference genes carrying an RNA-editing exception were also excluded from the primary matrix because a genomic stop or frame discrepancy could not be interpreted consistently across species. The resulting primary matrix contained 44 genes shared by all 18 species (792 coding sequences). Twelve sequences from six genes required a local single-exon boundary refinement guided by the reference protein. A sensitivity set removed all six genes, rather than treating the refined sequences as unqualified equivalents. Short genes remained in the complete matrix but were evaluated with minimum-length filters.

Reconstruction was checked against independent annotation where possible. All 44 primary *P. bifurcatum* sequences exactly matched MN623367.1. For *P. wallichii*, 42 primary orthologues were available in independently annotated NC_057599.1: 39 were exact matches to the reconstructed OR601555.1 sequences, and the remaining three had 99.46–99.92% aligned nucleotide identity. The two unavailable direct orthologues were not used to claim validation. These comparisons assess sequence and boundary concordance but do not convert the other records into independently annotated genomes.

Species-level codon profiles and phylogenetic comparisons

Terminal stop codons were removed before codon summaries. For each species, codons from the 44-gene primary matrix were pooled and RSCU was calculated for the 59 sense codons belonging to synonymous families (Sharp and Li 1986). Euclidean distances between the 59-dimensional RSCU vectors were divided by the square root of 59; this normalization changes the numerical scale but not the

ordering of pairwise distances. Species were also summarized by overall GC, GC at codon positions, GC at synonymous third positions, ENC, and parity-rule-2 coordinates.

The published branch-length plastome tree accompanying the public alignment was pruned to the same 18 taxa. An RSCU neighbour-joining phenogram was built from the normalized distances (Saitou and Nei 1987), and node support was estimated from 1,000 bootstrap replicates that resampled genes with replacement before recalculating species RSCU profiles. The phenogram was used as a multivariate summary rather than as an alternative species phylogeny. Topological overlap with the reference tree was described by the unrooted Robinson–Foulds distance and the number of shared non-trivial splits (Robinson and Foulds 1981).

Association between RSCU and reference-tree patristic distances was tested by a two-sided Spearman Mantel test with 9,999 label permutations (Mantel 1967). A restricted analysis retained only the 48 within-lineage species pairs and permuted labels separately within the Afro-American, Javan-Australian, and Malayan-Asian groups. This tested whether the relationship persisted after the deepest among-lineage contrasts were removed. Permutational multivariate analysis of variance provided a descriptive estimate of the fraction of RSCU-distance variation associated with the three published lineages. Principal-component analysis of the 59 codons supplied an ordination view of the same profiles.

Compositional and gene-level codon diagnostics

For each primary gene, species means were calculated for ENC, GC3, and GC12. Wright's ENC expectation under a composition-only model was plotted against GC3 (Wright 1990). A GC12–GC3 regression was used as a descriptive neutrality diagnostic, and parity-rule-2 coordinates summarized A-versus-T and G-versus-C asymmetry at fourfold-degenerate third positions. These plots were treated as diagnostic patterns rather than direct measurements of the proportion of codon variation caused by selection.

Mean pairwise gene RSCU divergence was compared with reference coding-sequence length and mean pairwise nucleotide p-distance using Spearman correlations. Phylogenetic analyses were repeated with all 84 stop-filtered genes and with primary genes of at least 300 or 600 nt. The all-gene analysis was interpreted as a sensitivity check because the additional loci included reference RNA-editing exceptions; it was not considered biologically superior to the conservative 44-gene matrix.

Exact-codon conditional randomization

Terminal stop codons were removed consistently before analysis; the first codon was fixed. Within each coding sequence, positions were grouped by their encoded amino acid. The native multiset of synonymous codons within each amino-acid group was shuffled among those positions. Consequently, every randomized sequence retained the amino-acid at every position, the exact count of every codon,

sequence length, total nucleotide composition, position-specific GC composition, and all within-codon 1→2 and 2→3 dinucleotide counts. Only synonymous-codon placement, codon pairs, and codon-boundary 3|1 dinucleotides could change.

Each species–gene sequence was randomized 9,999 times using a reproducible seed derived from a global seed (20260723) and a stable hash of species accession and gene. The implementation stopped if translation, codon multiset, length, nucleotide counts, start codon, or within-codon dinucleotides changed. All 792 sequences passed every invariant check. Random permutation P values included the native tail count plus one in the numerator and the number of randomizations plus one in the denominator, avoiding zero estimates (Phipson and Smyth 2010).

Boundary dinucleotides and effect sizes

All 16 DNA dinucleotides (AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, and TT) were counted at codon boundaries. For each species–gene sequence and dinucleotide, the primary effect was the native count minus its randomization mean, normalized per 100 codon boundaries. A negative value denotes depletion conditional on protein sequence and exact codon inventory. For the genus-wide descriptive landscape, native counts and null counts were summed across sequences within each permutation, and an observed/conditional-expectation ratio was calculated. CpG and TpA were specified before inspection as lower-tail targets; the other 14 dinucleotides supplied the complete background. Benjamini–Hochberg correction was applied to the 16 global lower-tail tests (Benjamini and Hochberg 1995).

Biological-level inference and sensitivity analyses

Permutation draws were treated as a null distribution, not as biological replicates. For CpG and TpA, effects were summarized separately across the 18 species and 44 genes. The principal uncertainty interval was a two-way hierarchical bootstrap: species rows and gene columns of the complete effect matrix were independently sampled with replacement, and the median of the crossed resample was calculated for 10,000 replicates. One-sided exact sign tests assessed whether species or gene median effects were more often negative than expected, with the two target dinucleotides adjusted as one family. Leave-one-species-out and leave-one-gene-out ranges evaluated concentration in individual units.

Sensitivity analyses used genes of at least 180, 300, 600, and 900 nt; excluded every gene with any boundary-refined sequence; and combined the 180-nt and no-refinement restrictions. Effects were also summarized within the three published lineages. Concordant directions among these related lineages were interpreted as retention across sampled groups, not as independent evolutionary origins. Analyses were performed with purpose-written JavaScript and Python scripts. GPT-5.6 assisted with drafting analysis code and initial language editing.

Results

Dataset audit and conserved compositional background

The 18 alignment rows were exact sequence matches to their stated GenBank accessions. Of 1,512 species–gene reconstructions, only three *rpl32* combinations failed the frame and stop audit. After the predefined editing-exception and integrity exclusions, the primary matrix contained 792 sequences from 44 genes. Twelve sequences (1.5%) required boundary refinement. Excluding all six genes affected by refinement left 38 genes and 684 species–gene sequences. The direct-annotation checks showed exact recovery of all 44 reference sequences and high concordance with the independently annotated *P. wallichii* plastome.

Species-level composition varied within narrow ranges. Overall GC ranged from 0.4090 to 0.4128, GC at synonymous third positions from 0.2949 to 0.3036, and ENC from 51.68 to 52.34. Thirty codons had RSCU greater than one in every species, and most ended in A or T. The three lineages therefore shared a strongly conserved plastid compositional background, but this background did not eliminate smaller differences among their 59-codon RSCU profiles.

Codon-use divergence tracked plastome phylogeny

Pairwise RSCU distance increased with reference-tree patristic distance (Mantel Spearman $\rho = 0.6685$, permutation $P = 0.0001$; Fig. 1d). The relationship was not explained only by contrasts among the three deepest groups: after between-lineage pairs were removed and labels were permuted within lineages, the 48 retained pairs gave $\rho = 0.8217$ and $P = 0.0006$. Descriptive within-lineage correlations were positive in the Afro-American, Javan-Australian, and Malayan-Asian groups, although the individual groups contained too few species for stable separate inference.

The RSCU neighbour-joining phenogram recovered the Afro-American, Javan-Australian, and Malayan-Asian species as three monophyletic clusters with gene-bootstrap support of 98.6%, 100.0%, and 72.5%, respectively (Fig. 1a, b). Ten of 15 non-trivial splits were shared with the reference plastome tree, giving a normalized Robinson–Foulds distance of 0.333. The phenogram therefore captured the major evolutionary partitions without reproducing every sequence-tree relationship. The first two RSCU principal components explained 26.2% and 22.5% of variance and separated the same broad lineages (Fig. 1c). Lineage membership accounted for 49.7% of RSCU-distance variation in the permutational multivariate analysis (pseudo-$F = 7.42$, $P = 0.0001$).

The agreement among the Mantel test, ordination, lineage-level multivariate analysis, and gene-bootstrap phenogram indicates that the result was not confined to a single summary statistic. At the same time, discordance among several internal branches shows that codon use did not simply reproduce the full plastome tree. Instead, the major lineages shared inherited compositional structure while individual genes and synonymous families accumulated partly distinct changes.

The phylogenetic association was stable to gene-set choice. Mantel ρ values were 0.7787 for all 84 stop-filtered genes, 0.6807 for the 24 primary genes of at least 300 nt, and 0.6601 for the 14 genes of at least 600 nt; all permutation P values were 0.0005. The higher coefficient in the all-gene matrix did not resolve the RNA-editing uncertainty but showed that the genus-wide relationship was not created by the conservative exclusion alone.

Composition and coding-sequence length shaped gene-level patterns

Most gene means fell below Wright's ENC–GC3 composition-only expectation (Fig. 2a), showing more unequal synonymous-codon use than predicted from GC3 alone. The GC12–GC3 regression was shallow and non-significant (slope = 0.153, R^2 = 0.034, P = 0.232; Fig. 2b). Parity-rule-2 coordinates were displaced from 0.5, with species values of A3/(A3 + T3) from 0.446 to 0.456 and G3/(G3 + C3) from 0.402 to 0.422 (Fig. 2c). These diagnostics indicate that third-position composition was asymmetric and that no single composition-only model explained all codon variation.

Gene-level RSCU divergence was unrelated to mean nucleotide p-distance (ρ = 0.2368, P = 0.122) but declined strongly with reference coding-sequence length (ρ = −0.6142, P = 9.23×10^{-6}; Fig. 2d). The largest apparent RSCU distances occurred mainly in short genes such as *psbK*, *psaJ*, *rpl33*, *psbN*, and *petL*. This result did not show that short genes lack evolutionary change; it showed that their ratio-based divergence estimates are more sensitive to a small number of codon substitutions.

Exact-codon randomization revealed a broad boundary landscape

Exact-codon randomization showed that codon boundaries were not exchangeable even after protein sequence and codon inventory were fixed (Fig. 3b; Electronic Supplementary Material 1). Nine dinucleotides had observed/expected ratios below one and seven had ratios above one. GpT was the most depleted boundary (ratio 0.865), followed by CpG (0.887), ApC (0.923), GpC (0.930), and TpA (0.933). The strongest enrichments were GpG (1.128), CpA (1.107), CpC (1.091), and TpT (1.091). The two prespecified targets therefore belonged to a broader redistribution of codon-boundary contexts; CpG was prominent but not uniquely depleted.

CpG depletion was consistent across species and genes

The median species–gene CpG effect was −0.517 occurrences per 100 codon boundaries (two-way bootstrap 95% confidence interval −0.860 to −0.236; Table 1). All 18 species had negative median CpG effects (one-sided sign P < 0.001 after correction), as did 34 of 44 genes (adjusted P < 0.001). Species medians remained negative in all three lineages: −0.650 in the Afro-American, −0.484 in the Javan-

Australian, and −0.445 in the Malayan-Asian group (Fig. 3c). Removing one species at a time gave medians from −0.528 to −0.499, and removing one gene at a time gave −0.547 to −0.483. The direction therefore did not depend on a single taxon or locus.

CpG depletion also persisted under stricter reconstruction and length rules (Fig. 3d). The median was −0.543 after all six boundary-refined genes were removed, −0.440 among the 30 genes of at least 180 nt, and −0.478 among the 25 genes satisfying both restrictions. Minimum lengths of 300, 600, and 900 nt produced medians of −0.468, −0.543, and −0.533, respectively. The proportion of negative species–gene effects rose from 0.715 in the full matrix to 0.907 at the 900-nt threshold.

TpA showed a pooled deficit but greater gene heterogeneity

TpA had a low genus-wide observed/expected ratio (0.933), and its median species–gene effect was −0.582 per 100 boundaries. All species medians were negative and the direction remained negative in the three lineages and all reconstruction and length sensitivities. Nevertheless, its two-way bootstrap interval extended slightly above zero (−0.947 to 0.024), and only 26 of 44 gene medians were negative (adjusted sign P = 0.146). Leave-one-out medians were uniformly negative, indicating that the uncertainty reflected variation distributed among genes rather than one influential locus. TpA was therefore supported as a pooled and species-level pattern but not as a gene-consistent effect of the same strength as CpG.

Table 1 Primary and sensitivity results for exact-codon conditional boundary effects.

Analysis	Biological units	Estimate	95% CI	Consistency	Interpretation
CpG, primary	18 species; 44 genes	Ratio 0.887; Δ −0.517	−0.860 to −0.236	18/18 species; 34/44 genes negative	Robust codon-order depletion
TpA, primary	18 species; 44 genes	Ratio 0.933; Δ −0.582	−0.947 to 0.024	18/18 species; 26/44 genes negative	Pooled/species signal; gene-heterogeneous
CpG, no refined genes	18 species; 38 genes	Δ −0.543	—	70.5% effects negative	Reconstruction sensitivity retained
CpG, ≥180 nt	18 species; 30 genes	Δ −0.440	—	72.4% effects negative	Short-gene sensitivity retained
CpG, ≥180 nt + no refined	18 species; 25 genes	Δ −0.478	—	71.3% effects negative	Combined sensitivity retained
CpG, leave one species out	18 analyses	Δ −0.528 to −0.499	—	All negative	Not driven by one species
CpG, leave one gene out	44 analyses	Δ −0.547 to −0.483	—	All negative	Not driven by one gene

Delta is the native count minus the randomization mean per 100 codon boundaries. CI, two-way species-by-gene bootstrap confidence interval. Sign-test P values are one-sided and corrected across the two prespecified dinucleotides.

Discussion

This study connects two scales of synonymous evolution in *Platycerium* plastomes. At the species level, divergence among 59-codon RSCU profiles tracked patristic distance and recovered the three established geographic lineages. At the local sequence level, codon-boundary CpG remained depleted after protein sequence and the exact inventory of every codon were fixed. These results are not alternative explanations of the same statistic. The RSCU analysis describes how synonymous composition differs among lineages, whereas the conditional randomization tests how already-used synonymous codons are ordered within genes. Their coexistence suggests that plastid codon evolution includes both a history-sensitive component and a more conserved sequence-context component.

The phylogenetic association was substantial despite narrow ranges of GC, GC3s, and ENC. This shows that evolutionary information can persist in the multivariate arrangement of individual synonymous codons even when aggregate composition changes little. The RSCU phenogram should not be interpreted as a replacement for the plastome sequence tree: it summarized a trait that is partly inherited and partly gene-dependent. Its recovery of the three major lineages, together with a positive restricted within-lineage Mantel relationship, is more appropriately read as evidence that synonymous differences

accumulated along both deeper and more recent branches. The incomplete match among internal splits is expected because RSCU distances compress information across genes and can be influenced by composition, gene length, and the particular synonymous families represented.

The shared preference for predominantly A- or T-ending codons is consistent with the A+T-rich substitution background common in many plastid genomes (Suzuki and Morton 2016). Such compositional conservation does not imply evolutionary stasis. A common background can persist from a shared plastid ancestor while lineage-specific substitutions gradually shift the relative use of individual codons. In this sense, the major lineages occupy different positions within a constrained compositional space rather than evolving entirely different codon systems. The pattern is compatible with a balance among mutation bias, drift, and lineage-specific functional or translational constraints.

Viewed as a multivariate trait, an RSCU vector is an emergent property of many synonymous substitutions distributed across the plastome. Its association with patristic distance suggests gradual divergence along the plastome phylogeny rather than an abrupt replacement of one codon regime by another. The conserved A/T-rich background may have been inherited through long-term organelle evolution, yet synonymous sites continued to turn over within that framework (Wicke et al. 2011; Daniell et al. 2016). The combination of narrow genome-wide GC ranges, lineage-associated RSCU differences, and heterogeneous gene-level distances therefore supports a nested model: older constraints delimit the available compositional space, lineage history shifts codon frequencies within that space, and individual genes respond according to length, amino-acid composition, functional context, and local sequence requirements. The present analyses cannot assign a rate to each layer, but they show that apparent compositional stability and ongoing synonymous evolution are compatible rather than contradictory.

The ENC–GC3, neutrality, and parity-rule-2 diagnostics add gene-level context but require restraint. Most genes fell below Wright's composition-only ENC expectation, the GC12–GC3 relationship was weak, and third-position parity was asymmetric. Together, these observations are compatible with influences beyond a single GC3-driven model, including gene-specific mutation spectra, selection on translational efficiency or accuracy, RNA-level constraints, and uneven amino-acid composition. They do not quantify the percentage of codon bias caused by natural selection. Plastid genes differ in expression and functional role, and plastid codon adaptation can vary among genes (Suzuki and Morton 2016), but the present public sequence data do not contain expression, tRNA abundance, or translational-efficiency measurements needed to identify the responsible process.

Coding-sequence length was a particularly strong source of gene-level heterogeneity. RSCU is a ratio within synonymous families, so short loci can move far in codon space after only a few substitutions. The negative relationship between gene length and apparent RSCU divergence therefore provides a statistical explanation for why several small photosynthetic or ribosomal genes appeared unusually divergent. It does not make those genes evolutionarily uninteresting; instead, it identifies them as hypotheses that require more species, within-species sampling, or count-based models. The persistence

of the species-level Mantel association after 300- and 600-nt filtering indicates that the broader phylogenetic result was not produced solely by these short-gene extremes.

The principal local result is the CpG deficit specifically at codon boundaries after conditioning on protein sequence and exact codon inventory. This null model removes several explanations that confound raw dinucleotide or codon-pair statistics: amino-acid composition, synonymous-codon abundance, overall and position-specific nucleotide composition, and within-codon dinucleotides are identical in native and randomized sequences. The remaining contrast concerns how already-used synonymous codons are distributed among compatible amino-acid positions. Exact-codon conditioning therefore identifies information in codon order that is invisible to total codon counts.

The all-16 analysis is essential for evolutionary interpretation. CpG was the second-most depleted boundary by observed/expected ratio, but GpT was lower and several other contexts were depleted or enriched. The evidence consequently supports a broader redistribution of neighbouring codon contexts rather than an isolated process targeting only CpG or TpA. Earlier studies described non-random CpG and TpA occurrence in plants and context-dependent mutation in chloroplast DNA (Boudraa and Perrin 1987; Morton 2003, 2023). The present result narrows that observation by showing that the boundary pattern cannot be explained by codon inventory alone. Possible contributors include context-dependent mutation or repair, RNA stability and structure, avoidance or creation of short sequence motifs, and interactions between translation and neighbouring codons.

The boundary analysis adds an evolutionary dimension that is not contained in ordinary codon-frequency summaries. Two sequences can encode the same protein and use exactly the same numbers of synonymous codons while arranging those codons differently. Persistent depletion under the conditional null therefore identifies a property on which mutation, repair, RNA-level constraint, or selection on neighbouring codons could act without changing amino-acid composition or total codon use. The negative CpG direction across all sampled species is compatible with a conserved ancestral state, repeated maintenance under comparable plastid constraints, or a mixture of inheritance and lineage-specific reinforcement. These alternatives make different phylogenetic predictions and could be separated in future by estimating gene-level boundary effects in denser fern sampling, reconstructing ancestral states, and comparing models of neutral drift, stabilizing constraint, and lineage-specific shifts.

CpG effects were negative in all sampled species, most genes, all three lineages, and every reconstruction or length subset. Because the lineages share plastid ancestry, these concordant directions are not three independent evolutionary origins. A parsimonious interpretation is that an ancestral sequence-context tendency was retained during diversification and then modified by gene-specific and lineage-specific substitutions. Repeated maintenance under similar plastid constraints is also possible. Distinguishing inheritance from convergent maintenance would require a denser fern phylogeny with outgroups and ancestral-state modelling of gene-level boundary effects.

TpA provides an informative contrast. Summed native counts were low, every species median was negative, and all sensitivity medians remained negative. Yet resampling genes widened the confidence

interval across zero and the gene-level sign test was not significant. This distributed heterogeneity would have been obscured if 9,999 randomizations or all 792 coding sequences had been treated as independent biological evidence. Under the two-way species-by-gene analysis, CpG is the more stable genus-wide feature, whereas TpA remains a plausible pooled and species-level tendency whose expression depends more strongly on gene context.

Several design features reduce the immediate reconstruction concerns. Public alignment rows were exact matches to the reported accessions, the annotated reference was recovered exactly, an independent *P. wallichii* plastome showed near-complete sequence concordance, and exclusion of every gene ever requiring boundary refinement preserved the result. Nonetheless, homology-guided reconstruction is not equivalent to independently curated annotation for every species. The primary matrix also excludes 39 reference genes with RNA-editing exceptions and cannot be generalized to their mature transcripts. One plastome per species does not measure within-species polymorphism, and lineage correlations cannot by themselves establish adaptive causality.

The integrated evolutionary picture is therefore one of constrained diversification. *Platycerium* plastomes share an ancestral A/T-rich compositional regime, their relative codon profiles have diverged in association with phylogenetic distance, and their coding sequences retain a conserved preference for particular codon-boundary contexts. Gene length and gene identity modulate how strongly these patterns are observed. Extending the analysis to independently annotated fern plastomes, population samples, outgroups, and matched transcript, tRNA, and expression data could test when the boundary organization arose and whether particular plastid functions or environments repeatedly favour the same codon contexts.

Declarations

Acknowledgements

The authors thank Mr. Lee, Jen-Chung Yang, Lai-You Yang, and Kuei-Ying Hsu for providing inspiration for this independent research. The authors also thank Szu-En Yang and Shun-Yuan Chang for providing equipment and a supportive environment in which to conduct the study.

Electronic supplementary material

Electronic Supplementary Material 1. *Platycerium* codon-evolution and conditional codon-order source data. The Excel workbook contains accession verification, reconstruction quality control, independent annotation validation, species codon metrics, RSCU ordination, phylogenetic association and sensitivity results, gene-level codon diagnostics, all 16 boundary-dinucleotide results, species- and gene-level effects, complete CpG and TpA coding-sequence effects, and metric definitions.

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing interests

The authors declare no financial, personal, or professional relationships that could have influenced the work.

Author contributions

Jun-Wei Yang conceived the study, curated the data, designed and performed the analyses, verified the scripts and results, interpreted the findings, prepared the figures, and drafted the manuscript. Cheng-Hsun Lee and Hsuan-Wei Huang contributed to study discussion, supervision, interpretation, and manuscript preparation and revision. Wei-Mei Hsieh and Su-Lan Hsu contributed to study discussion, interpretation, research-environment support, and critical review of the manuscript. All authors read and approved the final manuscript and agree to be accountable for the work.

Ethics approval

Not applicable. The study used only publicly available plant nucleotide sequences and involved no humans, animals, or newly collected biological material.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Data availability

Primary plastome sequences are available from GenBank under OR601544.1–OR601558.1, OR601560.1–OR601561.1, and MN623367.1. The published whole-plastome alignment and branch-length tree are Supporting Information of Ciarle et al. (2025), https://doi.org/10.1371/journal.pone.0329099. Accession validation, coding-sequence quality control, species codon metrics, RSCU ordination and distance results, phylogenetic sensitivity analyses, gene-level diagnostics, all 16 boundary dinucleotides, biological-level summaries, and reconstruction sensitivities are provided in Electronic Supplementary Material 1.

Code availability

Analysis scripts, configuration, and random seeds are retained in the author archive and can be supplied to the editor and reviewers for verification on request.

References

1. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B: Statistical Methodology 57:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
2. Boudraa M, Perrin P (1987) CpG and TpA frequencies in the plant system. Nucleic Acids Research 15:5729–5737. https://doi.org/10.1093/nar/15.14.5729
3. Ciarle R, de Bock K, Burns KC (2025) Evolutionary origins and life-history correlates of coloniality in the epiphytic fern genus *Platycerium* (Polypodiaceae). PLOS ONE 20:e0329099. https://doi.org/10.1371/journal.pone.0329099
4. Daniell H, Lin CS, Yu M, Chang WJ (2016) Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biology 17:134. https://doi.org/10.1186/s13059-016-1004-2
5. Kunec D, Osterrieder N (2016) Codon pair bias is a direct consequence of dinucleotide bias. Cell Reports 14:55–67. https://doi.org/10.1016/j.celrep.2015.12.011
6. Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Research 27:209–220.
7. Morton BR (2003) The role of context-dependent mutations in generating compositional and codon usage bias in grass chloroplast DNA. Journal of Molecular Evolution 56:616–629. https://doi.org/10.1007/s00239-002-2430-1
8. Morton BR (2023) Context and mutation in gymnosperm chloroplast DNA. Genes 14:1492. https://doi.org/10.3390/genes14071492
9. Phipson B, Smyth GK (2010) Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. Statistical Applications in Genetics and Molecular Biology 9:Article 39. https://doi.org/10.2202/1544-6115.1585
10. Robinson DF, Foulds LR (1981) Comparison of phylogenetic trees. Mathematical Biosciences 53:131–147. https://doi.org/10.1016/0025-5564(81)90043-2
11. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4:406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454
12. Sharp PM, Li WH (1986) An evolutionary perspective on synonymous codon usage in unicellular organisms. Journal of Molecular Evolution 24:28–38. https://doi.org/10.1007/BF02099948
13. Suzuki H, Morton BR (2016) Codon adaptation of plastid genes. PLOS ONE 11:e0154306. https://doi.org/10.1371/journal.pone.0154306
14. Tulloch F, Atkinson NJ, Evans DJ, Ryan MD, Simmonds P (2014) RNA virus attenuation by codon pair deoptimisation is an artefact of increases in CpG/UpA dinucleotide frequencies. eLife 3:e04531. https://doi.org/10.7554/eLife.04531
15. Wicke S, Schneeweiss GM, dePamphilis CW, Müller KF, Quandt D (2011) The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Molecular Biology

76:273–297. https://doi.org/10.1007/s11103-011-9762-4

16. Wright F (1990) The effective number of codons used in a gene. Gene 87:23–29. https://doi.org/10.1016/0378-1119(90)90491-9
17. Xue B, Huang E, Zhao G, Wei R, Song Z, Zhang X, Yao G (2024) 'Out of Africa' origin of the pantropical staghorn fern genus *Platycerium* (Polypodiaceae) supported by plastid phylogenomics and biogeographical analysis. Annals of Botany 133:697–710. https://doi.org/10.1093/aob/mcae003

Figures

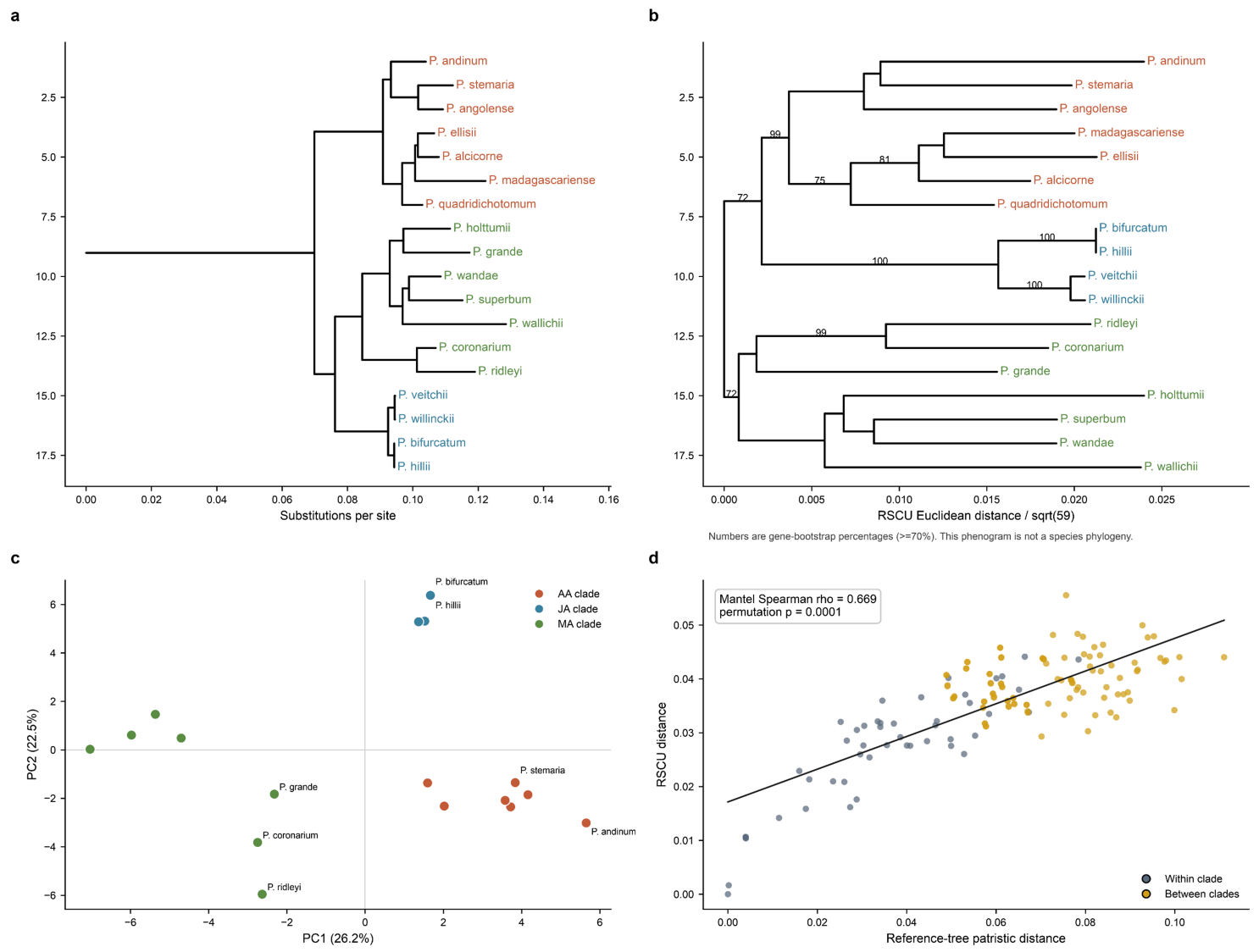


Figure 1

Phylogenetic and multivariate structure of plastid codon usage in 18 *Platycerium*species. (a) Published reference plastome tree pruned to the analysed taxa. (b) Neighbour-joining RSCU phenogram; labels are percentages from 1,000 gene bootstraps, and the phenogram is not a species phylogeny. (c) Principal-

component ordination of 59-codon RSCU profiles. (d) Pairwise RSCU distance versus reference-tree patristic distance. AA, Afro-American; JA, Javan-Australian; MA, Malayan-Asian

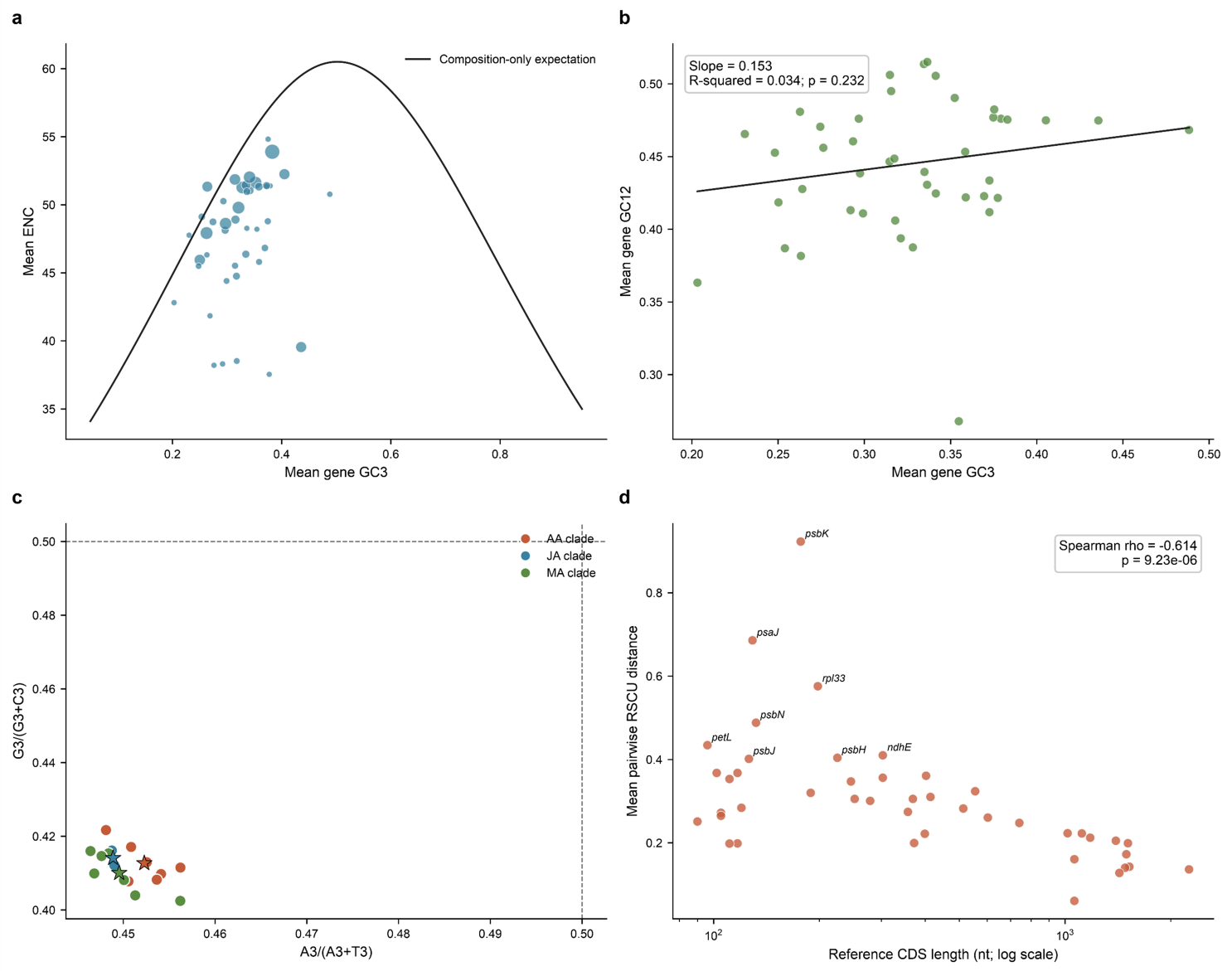


Figure 2

Compositional and gene-length diagnostics for the 44-gene primary matrix. (a) Mean gene ENC versus GC3; the solid curve is Wright's composition-only expectation, and point size represents reference coding-sequence length. (b) GC12-versus-GC3 neutrality diagnostic. (c) Parity-rule-2 plot for fourfold-degenerate families; stars mark lineage means and dashed lines indicate 0.5. (d) Reference coding-sequence length versus mean pairwise gene RSCU distance, showing that apparent gene-level divergence was greatest among short genes

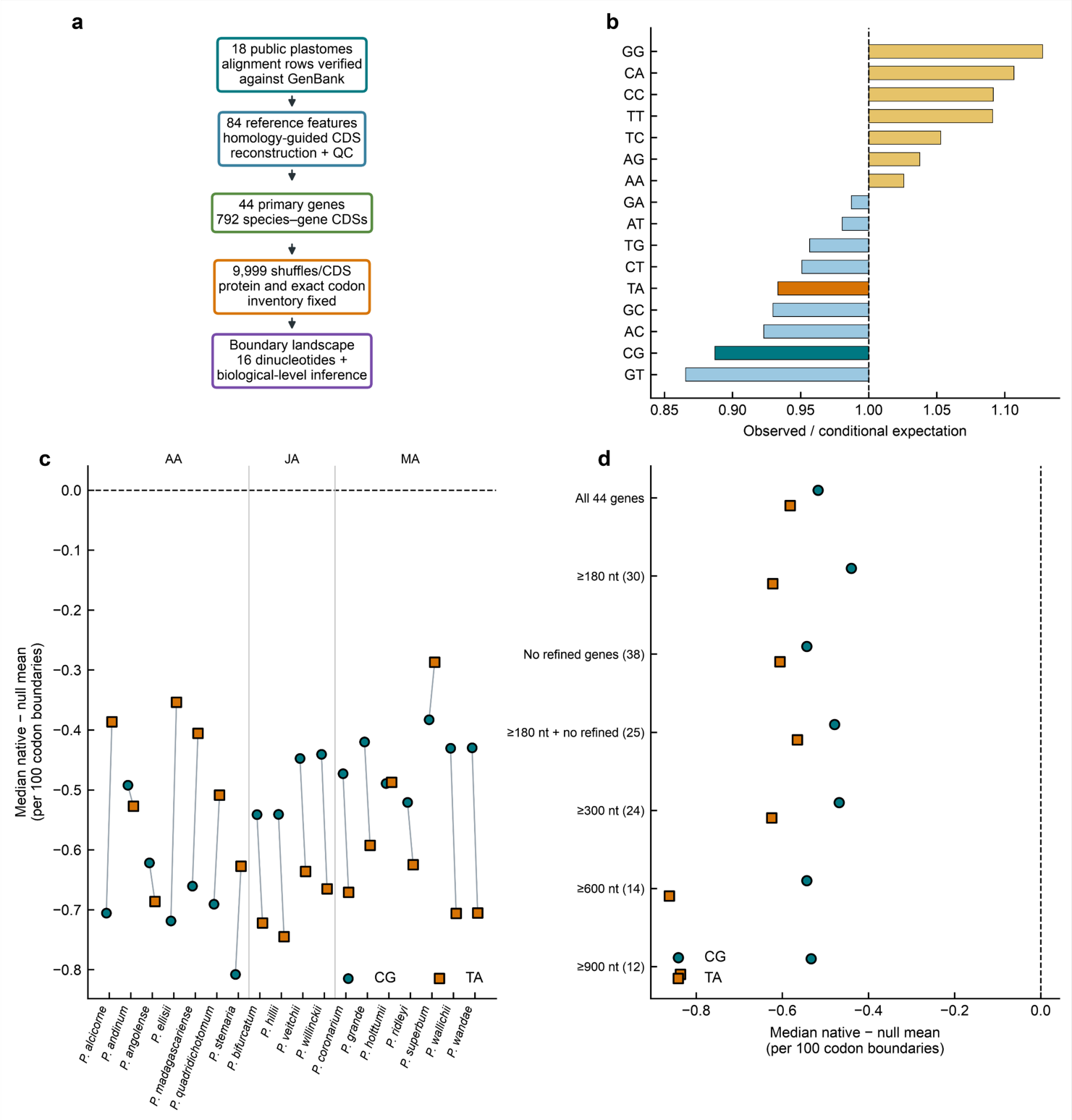


Figure 3

Conditional codon-order analysis of 18 *Platycerium* plastomes. (a) Analysis workflow. (b) Observed/conditional-expectation ratios for all 16 codon-boundary dinucleotides; CpG and TpA are distinguished from the remaining contexts by labelled bars, and the dashed line denotes equality. (c) Species median effects for CpG (circles) and TpA (squares), grouped by AA, JA, and MA lineages. (d)

Median effects under length and reconstruction sensitivities. Negative values indicate fewer native boundary occurrences than expected after conditioning on protein sequence and exact codon inventory

Supplementary Files

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

- 03ESM1PlatyceriumCodonEvolutionandConditionalOrderSourceData.xlsx