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.

Chloroplast genomes of *Vanilla* spp. (Orchidaceae) provide new evolutionary insights into the genus

Marileydi Zuta-Puscan
Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas https://orcid.org/0009-0006-0591-6888
Marco Pessoa-Filho
marco.pessoa@embrapa.br

Empresa Brasileira de Pesquisa Agropecuaria https://orcid.org/0000-0002-7333-4429
Alexander Damián-Parizaca
Universidad Tecnológica Perú https://orcid.org/0000-0002-0233-9935
Marília C. R. Pappas
Empresa Brasileira de Pesquisa Agropecuária https://orcid.org/0000-0002-4680-0199
Daniel Tineo
Instituto Nacional de Innovación Agraria (INIA) https://orcid.org/0000-0001-5263-6316
Jorge Ronny Díaz-Valderrama
Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas https://orcid.org/0000-0002-1015-3751
Angel David Hernández-Amasifuen
Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas https://orcid.org/0000-0001-8267-409X
Carla L. Saldaña
Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas https://orcid.org/0000-0002-9725-3578
José Jesús Tejada-Alvarado
Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas https://orcid.org/0000-0001-9989-1832
Juan Carlos Guerrero-Abad
Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas https://orcid.org/0000-0002-7285-9506
Manuel Oliva-Cruz
manuel.oliva@untrm.edu.pe

Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas https://orcid.org/0000-0002-9670-0970

Research Article

Keywords: comparative genomics, inverted repeat, Orchidaceae, plastid phylogenomics, structural rearrangement

Posted Date: August 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10727611/v1

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

Additional Declarations: The authors declare no competing interests.

Abstract

The genus *Vanilla* (Orchidaceae) holds significant economic value due to its production of aromatic compounds; however, numerous species are increasingly threatened by overexploitation and habitat destruction, underscoring the need for comprehensive genomic resources to support evolutionary and conservation efforts. In this study, we sequenced and assembled the complete chloroplast (cp) genomes of eight *Vanilla* species from Peru and Brazil. Additionally, two genomes were assembled from publicly available data. These were analyzed alongside seven chloroplast genomes from NCBI, yielding a dataset of 17 cp genomes. Genomic DNA extracted from young leaves was sequenced using the Illumina NovaSeq platform. Chloroplast genomes were assembled through a combination of *de novo* and reference-guided strategies and subsequently annotated using online platforms. The assembled cp genomes displayed the typical quadripartite structure, ranging from 147,372 to 163,306 bp in length and encoding 121–122 genes. Three highly variable regions ($\pi > 0.3$) and four genes (*clpP*, *rpl32*, *rps18*, and *infA*) with high nucleotide diversity were identified and proposed as promising molecular markers. Codon-based likelihood analyses provided convergent evidence of positive selection in five genes (*atpA*, *atpB*, *cemA*, *clpP*, and *ycf2*), supported by significant likelihood-ratio tests and high-probability Bayesian Empirical Bayes (BEB) sites. Phylogenetic analysis of concatenated coding sequences confirmed *Vanilla* as a monophyletic genus, with three well-supported clades corresponding to the current subgenus classification. These results provide important insights into the molecular evolution of *Vanilla* cp genomes, with relevant implications for phylogenetic, taxonomic, and conservation studies.

Main Conclusion

This study expands the genomic resources for Vanilloideae, increasing the number of complete chloroplast genomes in *Vanilla* from 7 to 17, including the first plastid genomes in Subgenus Membranacea.

Introduction

Vanilla, belonging to the orchid genus *Vanilla* Plum. ex Mill. is one of the most highly valued spices worldwide due to its wide-ranging applications in the culinary, cosmetic, and fragrance industries (Mahadeo et al. 2021). It represents the most diverse genus within the subfamily Vanilloideae, which includes the tribes Vanilleae and Pogonieae, and consists of approximately 128 species distributed throughout the tropical belt, with its greatest diversity occurring in the Neotropics (Karremans et al. 2025). *Vanilla* species have a pantropical distribution, broad ecological plasticity, evolutionary importance, and economic value. Notably, species such as *Vanilla planifolia* Andrews and *Vanilla pompona* Schiede are known for producing fruits with distinctive aromas and flavors, which confer significant commercial value (Havkin-Frenkel and Belanger 2018; Watteyn et al. 2023). However, increasing global demand and ongoing habitat fragmentation have resulted in several species being classified as Endangered (EN) by the International Union for Conservation of Nature (IUCN) (Bramel and Frey 2021).

Recent advances in *Vanilla* taxonomy, driven by comprehensive analyses of herbarium databases, type material, and newly collected specimens from targeted fieldwork, have substantially improved our understanding of species diversity within the genus. These efforts have facilitated the clarification of several previously poorly known taxa and have led to more precise species delimitations (e.g. Karremans et al. 2020; Damián-Parizaca et al. 2025a,b; Flanagan et al. 2025). Concurrently, the integration of DNA barcoding, phylogenetic inference, and high-throughput sequencing technologies has proven instrumental for clade recognition, species discrimination, and the reconstruction of evolutionary relationships within *Vanilla* (Hu et al. 2019; Besse et al. 2021; Chambers et al. 2021; Pansarin and Ferreira 2022; Damián-Parizaca et al. 2026; Karremans et al. 2025). As a result, previous studies support the monophyly of the genus *Vanilla*, revealing clades that reflect specific adaptations to pollination strategies and habitat conditions (Pansarin et al. 2012; Karremans et al. 2020). In the genus *Vanilla*, the most recent classification proposed by Karremans et al. (2025) recognizes four subgenera: *Gondwana* Karremans, Damian & Pupulin, *Membranacea* Karremans, Pupulin & Damian, *Tethyos* Karremans, Damian & Pupulin, and *Vanilla*, which are further subdivided into sections that reflect molecular clades and consistent morpho-ecological differences, with subg. *Vanilla*, including all species with vanillic and anisic aromatic compounds. This taxonomic framework integrates morphological variation and molecular evidence, offering a more comprehensive understanding of the group's complex evolutionary history.

In the genus *Vanilla*, as in most living plants (Archaeplastida), chloroplasts have a crucial role in various metabolic processes, including photosynthesis and the biosynthesis of amino acids and lipids (Gallage et al. 2018; Ko et al. 2020). Beyond these functions, recent studies suggest that chloroplasts are also involved in the biosynthesis of vanillin, the main aromatic compound found in subgenus *Vanilla*, through specialized metabolic pathways (Yang et al. 2017; Diamond et al. 2023).

Recent technological advances have enabled detailed analyses of chloroplast (cp) genomes, providing insights into their structure and composition (Guo et al. 2021). In most angiosperm plants, including the Orchidaceae family, cp genomes are typically organized into four distinct regions, separated by two inverted repeats (IR) sequences (Kim et al. 2023). These regions include a large single-copy (LSC) region and a small single-copy (SSC) region, which together encode between 110 and 130 genes vital for fundamental biological functions. From a genomic perspective, cp genomes show a highly conserved structural organization compared to nuclear genomes and are characterized by lower substitution rates and notable genetic stability (Wolfe et al. 1987). This stability, supported by maternal inheritance, low recombination rates, and the presence of hypervariable sites, enhances their usefulness as ideal markers for species identification, phylogenetic analysis, and population

genetics studies (Dong et al. 2012; Huotari and Korpelainen 2012). At the same time, chloroplasts undergo evolutionary processes such as deletions, insertions, substitutions, and rearrangements, providing key insights into evolution (Amiryousefi et al. 2018) and potential applications in genetic engineering (Bock and Khan 2004). Moreover, chloroplast diversification is influenced by the selective pressure from environmental factors during evolution (Sheikh-Assadi et al. 2022). Recent studies have found multiple genes under positive selection in various Orchidaceae genera, including *accD*, *matK*, *psbM*, *rpl20*, *rps12*, *ycf1*, and *ycf2* in *Paphiopedilum* (Liu et al. 2022); *ycf1a*, *rpoC2*, and *ycf2* in *Coelogyninae* (Li et al. 2024); as well as *ndhK*, *ycf2*, *ndhF*, *rbcL*, *rps4*, and *rps16* in *Bulbophyllum* (Tang et al. 2021), where the nonsynonymous substitution rate exceeds the synonymous rate. This indicates adaptive evolution related to photosynthetic efficiency, stress tolerance, and metabolism. Despite these findings, the understanding of positive selection patterns in *Vanilla* remains limited.

To date, only seven complete chloroplast genomes of *Vanilla* are available in the National Center for Biotechnology Information (NCBI), representing two subgenera, most of which belong to the subgenus *Vanilla*. In the highly cultivated *V. planifolia*, the plastid genome spans 148,011 base pairs (bp) and encodes 102 genes, including 30 tRNA genes and 4 rRNA genes. Its structural regions comprise an LSC of 86,358 bp, an SSC of 2,037 bp, and two IRs of 29,808 bp (Lin et al. 2015). In contrast, species of the subgenus *Tethyos*, such as *Vanilla aphylla* Blume, *Vanilla madagascariensis* Rolfe, and *Vanilla shenzhenica* Z.J. Liu & S.C. Chen, have cp genomes exceeding 150 kbp in total length, with notable variation in SSC and IR sizes (Kim et al. 2023; Li et al. 2019; Niu et al. 2017). These findings provide a strong foundation for more precise comparative and phylogenetic analyses, contributing to a deeper understanding of the evolutionary history of *Vanilla* spp.

In this study, we present the complete chloroplast genomes of ten *Vanilla* species, including seven species from the subgenus *Vanilla* and three from the subgenus *Membranacea*, for which chloroplast genome sequences are reported here for the first time. Using these data, we performed a comparative and phylogenetic analysis along with publicly available cp genomes, with the following objectives: (i) to characterize their structural organization, (ii) to identify hypervariable regions and repeats, (iii) to assess codon usage patterns and selective pressures, and (iv) to reconstruct the phylogenetic relationships within the genus. The findings offer new insights into cp genome evolution in *Vanilla* and provide a valuable framework for future research on taxonomy, conservation, and genetic improvement programs for orchids, which are highly prized in the global culinary industry.

Materials and methods

Material collection, DNA extraction, and sequencing

Plant material from Peru was collected from natural populations in the Peruvian departments of Loreto, San Martín, and Junín under a scientific research permit issued by the Servicio Nacional Forestal y de Fauna Silvestre (SERFOR; authorization code AUTIFL-2023-068) (Table 1). Species identification followed the taxonomic criteria proposed by Damian and Mitidieri (2020). Complete specimens (stems, leaves, and fruits when available) were pressed, dried, and deposited in recognized herbaria. Specimens of *Vanilla odorata* C. Presl, *Vanilla weberbaueriana* Kraenzl., and *Vanilla armoriquensis* Damián & Mitidieri were deposited at the KUELAP Herbarium of the Universidad Nacional Toribio Rodríguez de Mendoza in Amazonas, and *Vanilla acuta* Rolfe at the Herrense Herbarium of the Instituto de Investigaciones de la Amazonía Peruana (IIAP) in Loreto. Plant specimens from Brazil were collected in the States of Minas Gerais and Distrito Federal and are conserved *in vivo* at the Vanilla Active Germplasm Bank at Embrapa Genetic Resources and Biotechnology, Brasília, DF, Brazil. Samples were identified by Luciano de Bem Bianchetti and deposited at the CEN Herbarium, at Embrapa Genetic Resources and Biotechnology. They received the following accession numbers: BRA 00273076-0 (*Vanilla calyculata* Schltr.), BRA 00273063-8 (*Vanilla chamissonis* Klotzsch), BRA 00273080-2 (*Vanilla phaeantha* Rchb. f.), and BRA 00222045-7 (*V. pompona*).

Table 1. Collection sites of *Vanilla* species from Peru and Brazil

Species (Voucher)	Country	Department/ State	Altitude (m a.s.l.)	Coordinates	
				Lat	Lon
V. acuta (HH 7166)	Peru	Loreto	103	-3.83206287	-73.32123744
V. armoriquensis (KUELAP-3630)	Peru	Junín	1780	-11.04239005	-75.37049832
V. odorata (KUELAP-3647)	Peru	San Martín	810	-5.91429071	-77.11350320
V. weberbaueriana (KUELAP-3636)	Peru	San Martín	811	-5.91120912	-77.11362891
V. calyculata (CEN 113630)	Brazil	Minas Gerais	979	-15.53527800	-46.44805500
V. chamissonis (CEN 107433)	Brazil	Distrito Federal	1058	-15.73638900	-47.95139000
V. phaeantha (CEN 136764)	Brazil	Distrito Federal	1010	-15.72888900	-47.89833500
V. pompona (CEN00138828)	Brazil	Distrito Federal	1074	-15.63861100	-48.23027800

Before DNA extraction from Peruvian samples, young, healthy leaves were surface-sterilized with 70% ethanol, sectioned into 1 cm^2 fragments, and dried with silica gel. Genomic DNA was isolated from silica-dried leaves using a modified CTAB protocol (Doyle and Doyle 1987). Brazilian samples were freeze-dried and ground into a fine powder in a tissue homogenizer (Mini-BeadBeater-96, BioSpec Products, USA). The DNA extraction followed the protocol by Johnson et al. (2023), using an SDS lysis buffer and proteinase K. DNA quality and concentration were verified by electrophoresis on 1% TAE agarose gels and quantified with a Qubit fluorometer (Invitrogen, Carlsbad, CA, USA).

Paired-end sequencing libraries with an average insert size of 350 bp were constructed and prepared from fragmented genomic DNA. The libraries were sequenced using the Illumina NovaSeq platform, generating 150 bp paired-end reads. Sequencing services were provided by Macrogen Inc. (Seoul, South Korea) for Peruvian samples and by Novogene (Sacramento, CA, USA) for Brazilian samples.

Cp genome assembly and annotation

In addition to the eight species of *Vanilla* sequenced in this work, publicly available short-read sequencing data were downloaded from the Sequence Read Archive (SRA) at NCBI for the species *Vanilla mexicana* Mill. (SRR12628842) and *Vanilla x tahitensis* J.W. Moore (SRR12628843), for which no chloroplast genome assemblies were available. Raw reads for all samples were pre-processed with BBDuk from the BBtools suite (Bushnell 2014) for adapter trimming and cleaning of sequencing artifacts.

Processed reads were provided as input to GetOrganelle v1.7.7.1 (Jin et al. 2020) for *de novo* plastid genome assembly, using the 'embplant_pt' database, k-mer sizes of 21, 45, 65, 85, 105, and a word size of 102. When necessary, reads were reduced to achieve a target coverage of 2,000 or a maximum of 40 million reads. Assemblies were deemed successful if they were circularized. Specifically for *V. weberbaueriana* and *V. odorata*, the plastid genome was assembled using a hybrid strategy that combined *de novo* and reference-guided methods. Initially, quality-filtered reads were assembled separately with three *de novo* assemblers: MEGAHIT v1.2.9 (Li et al. 2015), GetOrganelle, and NOVOPlasty v4.3 (Dierckxsens et al. 2017), producing preliminary cp genome contigs for each sample. These assemblies were then integrated and refined through reference-guided mapping using the *V. planifolia* plastid genome (NC_026778.1) as a reference in Geneious Prime 2025.2 (Kearse et al. 2012). The "Map to Reference" function was performed with ten iterative cycles to improve assembly accuracy, increase coverage, and reduce potential errors.

Automated genome annotation was performed using GeSeq (Tillich et al. 2017), Geneious Prime, NCBI ORFfinder, and tRNAscan-SE v1.21(Lowe and Chan, 2016; http://lowelab.ucsc.edu/tRNAscan-SE/, accessed on May 20, 2025), followed by manual adjustments based on inspection of the resulting annotation files with Geneious Prime and Artemis v18.2.0 (Carver et al. 2012). Putative pseudogenes were identified based on disruption of open reading frames, including internal stop codons, frameshift mutations, truncated coding sequences, or absence of canonical start/stop codons relative to functional orthologs in closely related taxa (Yadav et al. 2023). Finally, cp genome maps were visualized with Chloroplot (Zheng et al. 2020; https://irscope.shinyapps.io/chloroplot/, accessed on September 15, 2025).

Cp genome structural features

The basic structural features of the newly sequenced *Vanilla* cp genomes, along with seven genomes retrieved from NCBI, were analyzed using Geneious Prime. Genome size, total gene content, and overall GC content were documented. Additionally, the GC content of the entire cp genome and the regions LSC, SSC, and IRs was calculated in RStudio v4.4.3 using the *seqinr* v.4.2-36 package (Charif and Lobry, 2007).

Comparative analyses of the IR boundaries were performed with IRscope (Amiryousefi et al. 2018; https://irscope.shinyapps.io/IRplus/, accessed on June 29, 2025). This tool enabled the visualization of junctions among the LSC, IR, and SSC regions, providing detailed insights into IR expansions, contractions, and structural stability across the analyzed cp genomes.

Analysis of repeat sequences

Simple sequence repeats (SSRs) were detected using the microsatellite identification tool (MISA) (Beier et al. 2017; https://webblast.ipk-gatersleben.de/misa/index.php?action=1). The following minimum repeat thresholds were applied: ≥10 repeat units for mononucleotide motifs, ≥5 for dinucleotides, ≥4 for trinucleotides, and ≥3 for tetra-, penta-, and hexanucleotide motifs. In addition, tandem repeats were detected using Tandem Repeats Finder (TRF) v. 4.09 (Benson 1999; https://tandem.bu.edu/trf/trf.html), employing default parameter settings.

Cp genome comparison and sequence divergence analyses

Comparisons of complete chloroplast genomes within the genus *Vanilla* were performed using the online Shuffle-LAGAN alignment program included in mVISTA, which detects rearrangements such as inversions (Frazer et al. 2004; https://genome.lbl.gov/vista/mvista/submit.shtml; accessed November 21, 2025), with *V. planifolia* NC_026778.1 designated as the reference sequence. Structural rearrangements across cp genomes were also assessed using Mauve v2.4.0 (Darling et al. 2004).

To identify divergent regions, the density of parsimony-informative sites was calculated across complete cp genomes and shared protein-coding genes using DnaSP v6.1 (Rozas et al. 2017). A sliding window approach was applied with a window size of 100 sites and a step size of 25 sites. Prior to this analysis, sequences were aligned with MAFFT v7.526 (Katoh and Standley 2013). Results were visualized in RStudio using the *ggplot2* package (Wickham 2016).

Codon usage analysis

Codon usage bias (CUB) was assessed using the relative synonymous codon usage (RSCU) index, calculated with CodonW v1.4.4 (http://codonw.sourceforge.net/, accessed 12 July 2025), which quantifies the preferential use of synonymous codons relative to the expectation of equal usage. In this context, RSCU values greater than 1 indicate codon preference, values lower than 1 reflect underrepresentation, and values equal to 1 suggest unbiased codon usage (Sharp and Li 1986). Codons with RSCU values above 1.6 were considered strongly overrepresented, whereas those with values below 0.6 were classified as underrepresented (Li et al. 2025a). Protein-coding genes (PCGs) were systematically extracted using PhyloSuite V1.2.2 (Zhang et al. 2020), with stringent filtering included, removing pseudogenes, excluding RNA-coding genes, and retaining sequences ≥300 bp to ensure statistical robustness across all species. Codon usage patterns were then illustrated using heatmaps generated with the ChiPlot online platform (https://www.chiplot.online/; accessed November 15, 2025).

CUB was further assessed using the effective number of codons (ENc), a common index that summarizes overall codon usage patterns, with values from 20, indicating extreme bias, to 61, reflecting uniform synonymous codon usage; genes with ENc values ≤35 were considered to show strong codon bias (Comeron and Aguadé 1998). ENc values were calculated using the *chips* tool implemented in the v. 6.6.0 EMBOSS package (Rice et al. 2000), and ENc-plot analysis (ENc vs. GC3S) was performed to identify the main forces influencing codon usage. In this analysis, genes clustering on or near the expected standard curve indicate codon bias driven mainly by mutational pressure, whereas deviations below the curve suggest the influence of natural selection or other evolutionary constraints (Wright 1990). To further assess the balance between mutation and selection, a neutrality analysis was conducted by examining the relationship between the GC content at the first and second codon position (GC12) and the GC content at the third codon position (GC3); regression slopes approaching zero indicate strong selective constraints, whereas slopes approaching 1 reflect mutation-driven codon usage (Sueoka 1988). Additionally, parity rule 2 (PR2) plot analysis was used to explore base composition biases at the third codon position, using G3/(G3+C3) and A3/(A3+U3) ratios, where symmetrical distributions around the center point indicate mutational equilibrium and significant deviations suggest the impact of selective forces (Sueoka 1995). All graphical analyses were produced using the *ggplot2* package in RStudio.

Selective pressure analysis

Adaptive evolution is the process by which organisms acquire traits that enhance their fitness in specific environments. In this study, shared protein-coding chloroplast gene sequences, excluding stop codons, were aligned using MACSE v2.07 (Ranwez et al. 2018) and then concatenated. Nonsynonymous (Ka) and synonymous (Ks) substitution rates, as well as their ratio ($\omega = Ka/Ks$), were estimated with the Yang and Nielsen (YN) model in KaKs_Calculator v3.0 (Zhang 2022). Values of $\omega < 1$ indicate purifying selection that maintains protein function, whereas $\omega > 1$ suggests positive selection that promotes adaptive changes (Nei and Gojobori, 1986). Genes with $\omega \geq 5$ or non-available values (NA), associated with limited nonsynonymous sites, were excluded due to unreliable evolutionary signals (Wu et al. 2020). Results were visualized with the *ggplot2* package in RStudio.

To identify specific codons under positive selection, site-branch models were used in EasyCodeML (Gao et al. 2019), an interface for PAML v4.9 (Yang 1997). Before analysis, fully conserved amino acid sequences were removed. The remaining PCGs were realigned with MACSE v2.07 and used to build an unrooted maximum likelihood phylogenetic tree in IQ-TREE v.3.0.1(Wong et al. 2025), which served as input for evolutionary modeling. Site models in CodeML treat ω as a random variable across codons, allowing it to vary among sites (Álvarez-Carretero et al. 2023). Two model pairs were tested: M1a (nearly neutral) vs. M2a (positive selection), and M7 (beta) vs. M8 (beta + ω). In these comparisons, M1a and M7 assume no codons with $\omega > 1$ (null models), while M2a and M8 allow for codons under positive selection (alternative models). Likelihood ratio tests (LRT) were performed to compare null and alternative models (Yang et al. 2000), with p-values estimated under a chi-square distribution in PAML v4.9. Sites with posterior probability ≥0.95 under Bayesian Empirical Bayes (BEB) analysis were considered to be under positive selection (Yang et al. 2005).

Phylogenetic analyses

Coding sequence (CDS) regions from 24 complete Orchidaceae chloroplast genomes were exported from annotated plastids as FASTA files (one FASTA file per CDS per species) in Geneious Prime. Samples included 20 species of subfamily Vanilloideae, 18 species from tribe Vanilleae, and two from tribe Pogonieae. Within the *Vanilla* genus, three species belong to subgenus *Membranacea*, section *Membranacea*; four species belong to subgenus *Tethyos*, being one from section *Aphyllae* Rolfe and three from section *Miguelia* (Aver.) Karremans, Damián & Pupulin; and eight

species belong to subgenus *Vanilla*, section *Vanilla* (Karremans et al. 2025). *Cypripedium debile* Rchb. f. and *Phragmipedium besseae* Dodson & J.A. Kuhn (Orchidaceae, Cypripedioidea) were used as outgroups.

Gene FASTA files (including all samples with a given CDS) were generated using a bash script. Each FASTA was aligned with MAFFT using the '--localpair' and '--maxiterate 1000' options. The aligned FASTA files were concatenated into a single file using a custom Python script, adding gaps for samples lacking a given CDS. The concatenated FASTA was trimmed using ClipKIT v.2.4.1 (Steenwyk et al. 2020) and the dynamic gap threshold (smart-gap) method. ModelFinder (Kalyaanamoorthy et al. 2017) in IQ-TREE v.3.0.1 was used to identify a DNA substitution model for Maximum-Likelihood phylogenetic analysis, based on the Bayesian Information Criterion (BIC).

Phylogenetic trees were constructed using Maximum Likelihood (ML) and Bayesian Markov Chain Monte Carlo inference (BI). The ML analysis was conducted using RAxML Next Generation v1.2.2 (Stamatakis 2014; Flouri et al. 2015; Kozlov et al. 2019), with the GTR+F+I+R3 model. Analysis parameters included ML tree search with 200 random and 200 parsimony starting trees. Bootstrapping used Majority Rule Extended (MRE) based bootstrapping after every 200 replicates. Branch support metrics were computed with the Felsenstein bootstrap and the Transfer Bootstrap Expectation. The Bayesian analysis was performed with MrBayes v3.2.7a (Ronquist et al. 2012), using a mixed substitution model with a gamma distribution, default priors, and two independent runs with four chains each. The runs consisted of 1,000,000 generations, sampling every 500 generations. The first 5,000 generations were discarded as burn-in. Trees were visualized with FigTree v1.4.2.

Results

General features of *Vanilla* chloroplast genome

The cp genomes of *Vanilla* species assembled in this study were recovered as single circular DNA molecules (Figure 1), and the subgenus *Vanilla* cp genomes ranged from 147,372 bp (*V. phaeantha*) to 150,748 bp (*V. calyculata*). In contrast, species in the subgenus *Membranacea* exhibited notably larger cp genomes, ranging from 161,310 bp (*V. acuta*) to 163,306 bp (*V. armoriquensis*) (Table 2). All assembled cp genomes exhibited the typical quadripartite structure of angiosperms, consisting of a large single-copy (LSC) region, a small single-copy (SSC) region, and a pair of inverted repeats (IRs) (Figure 1). The sizes of the LSC, SSC, and IR regions ranged from 86,061 to 94,913 bp, 1,254 to 3,103 bp, and 29,669 to 32,657 bp, respectively, reflecting moderate structural variation between subgenera (Table 2). The total GC content of the plastomes ranged from 34.7% to 35.5%, where the inverted repeat (IR) region exhibited a significantly higher GC content (38.6% - 40.4%) than the SSC and LSC regions (23.8% - 27.7% and 31.7% - 33%, respectively) (Table 2).

Table 2. Basic characteristics of seventeen complete chloroplast genomes of the *Vanilla* genus

Taxa	GenBank ID	Length (bp) / GC content (%)				Gene content (unique)			
		Total	LSC	IR	SSC	Total	PCGs	tRNA	rRNA
Vanilla* Subgenus *Membranacea									
Section *Membranacea*									
V. acuta Rolfe	PX659496[a]	161,310/35.4	93,523/32.2	32,406/40.4	2,975/26.9	121(102)	75(68)	38(30)	8(4)
V. armoriquensis Damián & Mitidieri	PX659497[a]	163,306/35.1	94,913/31.7	32,645/40.4	3,103/27.2	121(102)	75(68)	38(30)	8(4)
V. mexicana Mill.	BK075389[b]	161,499/35.4	93,103/32.2	32,657/40.4	3,082/27.7	121(102)	75(68)	38(30)	8(4)
Vanilla* Subgenus *Tethyos									
Section *Aphyllae*									
V. madagascariensis Rolfe	NC_046809	151,552/34.7	87,490/31.9	31,404/38.7	1,254/23.1	121(101)	74(67)	39(30)	8(4)
Section *Miguelia*									
V. aphylla Blume	NC_035320	150,165/35.0	87,379/32.4	29,716/39.4	3,354/24.9	119(100)	72(66)	39(30)	8(4)
V. somae Hayata	NC_079955	150,005/35.3	86,439/32.7	30,901/39.3	1,764/23.9	122(102)	75(68)	39(30)	8(4)
V. shenzhenica Z.J.Liu & S.C.Chen	MK962478	151,537/35.0	87,487/32.4	31,564/38.6	922/27.9	117(99)	70(65)	39(30)	8(4)
Vanilla* Subgenus *Vanilla									
Section *Vanilla*									
V. weberbaueriana Kraenzl.	PX659495[a]	149,585/35.1	86,856/32.6	30,282/39.1	2,165/23.8	122(102)	75(68)	39(30)	8(4)
V. odorata C.Presl.	PV739282[a]	148,546/35.4	86,424/32.9	30,040/39.4	2,042/25.9	122(102)	75(68)	39(30)	8(4)
V. phaeantha Rchb.f.	PX058851[a]	147,372/35.5	86,071/33.0	29,669/39.5	1,963/25.7	122(102)	75(68)	39(30)	8(4)
V. × tahitensis J.W.Moore	BK075390[b]	148,018/35.4	86,365/32.9	29,808/39.5	2,037/24.7	122(102)	75(68)	39(30)	8(4)
V. pompona Schiede	PX058849[a]	148,821/35.3	86,474/32.8	30,119/39.3	2,109/25.1	122(102)	75(68)	39(30)	8(4)
V. chamissonis Klotzsch	PX058850[a]	150,408/35.2	87,607/32.6	30,323/39.2	2,155/24.5	122(102)	75(68)	39(30)	8(4)
V. calyculata Schltr.	PX058852[a]	150,748/35.2	87,981/32.7	30,300/39.2	2,167/24.5	122(102)	75(68)	39(30)	8(4)
V. planifolia Andrews	NC_026778	148,011/35.4	86,358/32.9	29,808/39.5	2,037/24.7	122(102)	75(68)	39(30)	8(4)
V. planifolia Andrews	MN200375	147,714/35.4	86,061/32.9	29,808/39.5	2,037/24.7	122(102)	75(68)	39(30)	8(4)
V. pompona Schiede	NC_036809	148,009/35.4	86,358/32.9	29,807/39.5	2,037/25.0	122(102)	75(68)	39(30)	8(4)

[a]Sequenced and assembled in the present study. [b]Assembled in this study from publicly available SRA data (*Vanilla mexicana*, SRR12628842; *Vanilla × tahitensis*, SRR12628843).

Across the three *Vanilla* subgenera, cp genomes encoded a mostly conserved gene set. Species of subgenus *Membranacea* contained 121 genes, those of subgenus *Tethyos* harbored 117 - 122 genes, and species of subgenus *Vanilla* consistently possessed 122 genes. Additionally, most protein-coding genes showed highly similar lengths across species, reflecting strong structural and functional conservation of the cp genome; however, a few genes, particularly *accD* and *ycf1*, displayed noticeable length variation (Table S1). Overall, the cp genomes included 70 - 75 PCGs, 38 - 39 tRNA genes, and eight rRNA genes (Table 2). Excluding redundancy introduced by inverted repeat (IR) duplication, gene content ranged from 99 to 102 genes among species. Gene content variation was limited exclusively to PCGs, with subgenera *Membranacea* and *Vanilla* each maintaining 68 PCGs, while species of subgenus *Tethyos* exhibited a slightly reduced and variable set of 65 - 68 PCGs (Table 2). Notably, a widespread loss of genes encoding NADH dehydrogenase subunits (*ndh* genes) was observed across all *Vanilla* plastomes. All *ndh* genes were absent, except for *ndhB*, which remained as a pseudogene in all species analyzed (Table 3; Figure 1). Pseudogenes were identified in three species of subgenus *Tethyos*: *V. madagascariensis* contained a pseudogenized *ccsA*, *V. aphylla* had pseudogenes of *accD* and *ycf1*, and *V.*

shenzhenica showed pseudogenization of *psbC*, *ccsA*, and *rps7*. These pseudogenes were primarily characterized by deletions, with larger deletions in *ccsA*, *accD*, and *ycf1*, whereas *psbC* and *rps7* exhibited single-nucleotide deletions (Table S1). Concerning tRNA genes, subgenus *Membranacea* uniquely contained *trnL-UAG* in the SSC region, whereas in subgenera *Tethyos* and *Vanilla* this gene was located within the IR regions (Table 3; Figure S1). This structural difference explains the slightly lower number of tRNA genes found in *Membranacea* cp genome (Table 2).

Table 3. Gene composition in *Vanilla* chloroplast genomes

Category of genes	Group of genes	Name of genes
Photosynthesis	Photosystem I	*psaA*, *psaB*, *psaC*, *psaI*, *psaJ*, *ycf3*[##], *ycf4*
	Photosystem II	*psbA*, *psbB*, *psbC*[a], *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbN*, *psbT*, *psbZ*
	Rubisco	*rbcL*
	ATP synthase	*atpA*, *atpB*, *atpE*, *atpF*[#], *atpH*, *atpI*
	cytochrome b/f complex	*petA*, *petB*[#], *petD*[#], *petG*, *petL*, *petN*
	Subunits of NADH-dehydrogenase	*ndhB*[Ψ]
RNA genes	Ribosomal RNAs	*rrn4.5*, rrn5*, rrn16*, rrn23**
	Transfer RNAs	*trnA-UGC**[#], *trnC-GCA*, *trnD-GUC*, *trnE-UUC*, *trnF-GAA*, *trnfM-CAU*, *trnG-GCC*[#], *trnG-UCC*, *trnH-GUG**, *trnI-CAU**, *trnI-GAU**[#], *trnK-UUU*[#], *trnL-CAA**, *trnL-UAA*[#], *trnL-UAG**[b], *trnM-CAU*, *trnN-GUU**, *trnP-UGG*, *trnQ-UUG*, *trnR-ACG**, *trnR-UCU*, *trnS-GCU*, *trnS-GGA*, *trnS-UGA*, *trnT-GGU*, *trnT-UGU*, *trnV-GAC**, *trnV-UAC*[#], *trnW-CCA*, *trnY-GUA*
Transcription and translation	Small ribosomal subunit	*rps2*, *rps3*, *rps4*, *rps7**[a], *rps8*, *rps11*, *rps12**[##], *rps14*, *rps15*, *rps16*[#], *rps18*, *rps19**[c]
	Large ribosomal subunit	*rpl2**[#], *rpl14*, *rpl16*[#], *rpl20*, *rpl22*, *rpl23**, *rpl32*, *rpl33*, *rpl36*
	DNA dependent RNA polymerase	*rpoA*, *rpoB*, *rpoC1*[#], *rpoC2*
Other genes	Subunit of Acetyl-CoA-carboxylase	*accD*[d]
	C-type cytochrom synthesis	*ccsA*[a, e]
	Envelop membrane protein	*cemA*
	Protease	*clpP*[##]
	Maturase	*matK*
	Translational initiation factor	*infA*
Unkown function	Conserved open reading frame	*ycf1**[d], *ycf2**

* Two gene copies in IRs; [#] genes containing one intron; [##] genes containing two introns; [Ψ] pseudogene in all species; [a] pseudogene in *V. shenzhenica*; [b] located in the SSC region in species of subgenus *Membranaceae*; [c] present in only one IR in *V. shenzhenica*; [d] pseudogene in *V. aphylla*; [e] pseudogene in *V. madagascariensis*.

Simple Sequence Repeat (SSR) and tandem repeat analyses

Simple sequence repeats (SSRs) representing all six motif classes examined (mono-, di-, tri-, tetra-, penta-, and hexanucleotide repeats) were identified across 17 chloroplast genomes representing 15 *Vanilla* species. A total of 2,153 SSRs were detected, although their abundance varied markedly among plastomes, ranging from 105 in *V. phaeantha* to 158 in *V. madagascariensis* (Figure 2A; Table S2). Mononucleotide SSRs were consistently the predominant repeat class, accounting for 983 loci across the dataset and ranging from 45 in *V. mexicana* to 69 in *V.*

weberbaueriana (Figure 2B). Tetranucleotide repeats were the second most abundant class (372 loci), followed by dinucleotide (348 loci) and trinucleotide repeats (255 loci). In contrast, penta- and hexanucleotide repeats were less frequent, accounting for 104 and 91 loci, respectively. At the individual plastome level, tetranucleotide repeats ranged from 16 to 29 loci, dinucleotide repeats from 15 to 26, trinucleotide repeats from 6 to 36, pentanucleotide repeats from 2 to 15, and hexanucleotide repeats from 2 to 12 (Figure 2; Table S2).

SSR distribution across the structural regions of the plastome was markedly heterogeneous. The large single-copy (LSC) region contained the greatest proportion of SSRs in all samples, accounting for 69.8% of the total SSR complement in *V. aphylla* and up to 80.3% in both *V. planifolia* accessions and *V. × tahitensis*. In contrast, the small single-copy (SSC) region contained only 0–5 SSRs per plastome. The IRB and IRA regions contained equivalent numbers of SSRs within each plastome, consistent with their duplicated organization as inverted repeats (Table S2).

Because mononucleotide SSRs represented the predominant repeat class, their composition was examined in greater detail. Of the 983 mononucleotide loci identified, 520 (52.9%) were T-based motifs and 448 (45.6%) were A-based motifs, together accounting for 98.5% of all mononucleotide SSRs (Figure 3; Table S2). In contrast, C- and G-based motifs were rare, with only 7 (0.7%) and 8 (0.8%) loci, respectively. The most frequent motifs were T(10), with 310 occurrences, and A(10), with 284, followed by T(11) and A(11), with 116 and 115 occurrences, respectively. Overall, motif abundance declined with increasing homopolymer length, whereas longer motifs displayed more restricted distributions among plastomes.

Hierarchical clustering based on Z-scores of mononucleotide motif abundance revealed distinct SSR profiles among the analyzed plastomes (Figure 3). A particularly consistent pattern was observed for *V. planifolia*: the two accessions clustered closely together and exhibited completely concordant mononucleotide SSR profiles, with 66 SSRs in each plastome and identical counts for all 18 motifs examined. This similarity was not restricted to mononucleotide SSRs, as both accessions also contained the same total number of SSRs (127) and displayed identical SSR distributions across the LSC, IRB, SSC, and IRA regions (Table S2). In contrast, the two accessions identified as *V. pompona* showed a different pattern. They did not cluster together and contained 68 and 58 mononucleotide SSRs, respectively. Their profiles differed in 10 of the 18 motifs examined, with the largest discrepancy observed for T(10), which occurred 25 times in *V. pompona*1 compared with 16 times in *V. pompona*2. The two accessions also differed in their overall SSR abundance, containing 130 and 114 loci, respectively. Collectively, these results show that although some conspecific plastomes can exhibit highly conserved SSR profiles, as observed in *V. planifolia*, this pattern was not uniform across the accessions examined, and the overall clustering based on mononucleotide SSRs did not strictly recapitulate the currently recognized infrageneric classification of *Vanilla*.

Regarding tandem repeats (TRs), a total of 4,592 repeats were identified across the 17 plastomes, with marked variation among samples (Table S3). The total number ranged from 198 TRs in *V. phaeantha* to 363 in *V. madagascariensis*, with comparatively high numbers also recorded in *V. armoriquensis* (347), *V. shenzhenica* (319), and *V. aphylla* (318). Repeat-unit lengths ranged from 7 to 206 bp, with a clear predominance of shorter repeats. The 7–15 bp and 15–30 bp classes comprised 1,970 (42.9%) and 2,116 (46.1%) repeats, respectively, together accounting for 89.0% of all TRs detected. In contrast, repeat frequency progressively decreased with increasing repeat-unit length; TRs longer than 75 bp were uncommon, and those exceeding 90 bp represented only 0.7% of the dataset. The longest repeat unit reached 206 bp and was detected in *V. madagascariensis*. TR distribution across plastome structural regions was also heterogeneous, with the highest abundance occurring in the LSC region and considerably fewer repeats occurring in the SSC region (Table S3).

Cp genome comparison and sequence divergence analyses

Global alignment of the 17 *Vanilla* chloroplast genomes using mVISTA, with *V. planifolia* (NC_026778.1) as the reference, revealed a high degree of overall sequence conservation, although clear differences were observed among the major infrageneric lineages (Figure 4). Plastomes belonging to subgenera *Tethyos* and *Membranacea* showed greater sequence divergence from the reference than those assigned to subgenus *Vanilla***.** These differences were distributed throughout the plastome and were particularly evident in several non-coding regions.

Most coding regions were highly conserved across the analyzed plastomes. Nevertheless, several genes, including *accD, ycf1,* and *ycf2*, displayed comparatively greater sequence variation. Divergence was more pronounced within intergenic regions, particularly in the intervals *trnK-UUU–rps16, rps16–trnQ, trnR–atpA, atpF–atpH, atpI–rps2, rpoB–trnC, trnT–psbD, atpB–rbcL, rbcL–accD, accD–psaI, clpP–psbB,* and *rps8–rpl14*, which showed noticeable reductions in sequence identity across several of the compared plastomes (Figure 4).

Comparison of the junctions between the inverted repeats and single-copy regions revealed variation in the positions of the LSC/IRB (JLB), IRB/SSC (JSB), SSC/IRA (JSA), and IRA/LSC (JLA) boundaries among *Vanilla* plastomes (Figure S1). The JLB boundary was located within or closely associated with *rpl22*, although the extent of this gene into the IRB region varied among species. The JLA boundary was generally positioned between *rps19***,** associated with IRA, and *psbA*, located in the LSC region. *Vanilla shenzhenica* exhibited a distinct configuration in which only *psbA* was associated with this junction. The most pronounced differences were observed at the boundaries involving the SSC region. In several plastomes, *ccsA* was associated with the JSB and JSA junctions, whereas species of subgenus *Membranacea* displayed a distinct boundary configuration. In this lineage, *trnL* was located near the SSC/IRA junction, distinguishing it from the patterns observed in subgenera

Vanilla and *Tethyos*. Variation in the relative positions of genes adjacent to the SSC/IR junctions was also observed among some species of *Tethyos*. Overall, the major differences in IR/SC boundary organization among the analyzed plastomes were concentrated around the SSC region.

Progressive alignment of the 17 *Vanilla* plastomes using Mauve identified four locally collinear blocks (LCBs) and revealed differences in genome organization among the three subgenera analyzed (Figure S2). Plastomes belonging to subgenera *Vanilla* and *Tethyos* generally retained the order and orientation of these blocks, although localized variation was observed, particularly in regions associated with the IR/SSC boundaries. In contrast, all three plastomes of subgenus *Membranacea* shared a conspicuous structural rearrangement within the LSC region, corresponding to an inversion of approximately 39 kb (Figures 5 and S2).

The inversion identified in *Membranacea* encompassed 29 protein-coding genes and 12 tRNA genes. Detailed comparison of gene order among representatives of the three subgenera delimited the inverted segment to regions located near *trnS-GCU* and *trnS-GGA* (Figure 5). The relative position and orientation of these genes differed between *Membranacea* and representatives of *Tethyos* and *Vanilla*, coinciding with the boundaries of the inverted segment. No SSRs or longer tandem repeats were detected within or immediately adjacent to these breakpoints based on the repeat analyses performed in this study (Tables S2 and S3).

Sliding-window analysis of nucleotide diversity revealed a heterogeneous distribution of sequence variation across the plastome alignment (Figure 6A). Values of π ranged from 0 to approximately 0.33, with the highest peaks concentrated in regions including the *ndhB* pseudogene and the *rps12–rps7* interval. Overall, non-coding regions displayed more frequent and pronounced peaks of nucleotide diversity than most coding regions. Independent analysis of the 63 shared protein-coding sequences showed that most genes exhibited relatively low levels of nucleotide diversity (Figure 6B). Nevertheless, four genes showed π values greater than 0.05. Among them, *clpP* displayed the highest nucleotide diversity, reaching approximately 0.12, followed by *rps18* and *rpl32*, both with values close to 0.07, whereas *infA* showed a value slightly above 0.05. These four genes represented the most variable protein-coding regions among the *Vanilla* plastomes analyzed.

Codon usage analysis

Relative synonymous codon usage (RSCU) was evaluated across all 64 codons of the genetic code in the 17 *Vanilla* plastomes. For graphical comparison of codon usage patterns, the heatmap included the 59 informative synonymous codons, excluding the three stop codons and the single-codon amino acids methionine (AUG) and tryptophan (UGG) (Figure 7; Table S5). RSCU profiles were highly consistent across plastomes, with limited variation among species and subgenera. Among the synonymous codons represented, RSCU values ranged from 0.289 to 1.918. Thirty codons had mean RSCU values > 1, of which 28 ended in A or U, revealing a pronounced preference for these nucleotides at the third codon position. The highest mean RSCU values were observed for UUA (Leu; RSCU = 1.892), GCU (Ala; 1.827), UCU (Ser; 1.800), AGA (Arg; 1.782), and ACU (Thr; 1.674). In contrast, CGC (Arg) showed the lowest mean RSCU value (0.305), followed by CUC (Leu; 0.367) and AGC (Ser; 0.394). Overall, codon preference patterns were broadly conserved across the three subgenera, as reflected by the highly similar RSCU profiles shown in the heatmap (Figure 7).

The effective number of codons (ENC) ranged from 36.19 to 61.34 across the 629 gene–plastome observations representing 37 genes from the 17 plastomes (Figure 8A; Table S6). ENC values ranged from 37.39 to 61.34 in subgenus *Membranacea*, from 38.75 to 57.22 in *Tethyos*, and from 36.19 to 56.83 in subgenus *Vanilla*. None of the genes showed ENC values ≤ 35. In the ENC–GC3s plots, approximately two-thirds of the observations were located below Wright's theoretical curve: 66.7% in *Membranacea*, 66.9% in *Tethyos*, and 68.6% in *Vanilla*. Overall, 427 of the 629 observations (67.9%) fell below the theoretical expectation (Figure 8A).

Neutrality analysis revealed positive relationships between GC3 and GC12 in all three subgenera, although both the regression slopes and coefficients of determination were relatively low (Figure 8B; Table S7). In *Membranacea*, the relationship was described by $y = 0.376 + 0.211x$, with $R^2 = 0.028$. In *Tethyos*, the regression slope increased to 0.343 ($y = 0.332 + 0.343x$; $R^2 = 0.064$), whereas subgenus *Vanilla* showed the steepest slope, at 0.456 ($y = 0.300 + 0.456x$; $R^2 = 0.095$). Despite these differences among subgenera, GC3 accounted for less than 10% of the variation in GC12 in all three groups.

The PR2 analysis further revealed an asymmetric distribution of nucleotides at the third codon position (Figure 8C; Table S8). Quadrant IV [G3/(G3+C3) > 0.5 and A3/(A3+T3) < 0.5] contained the largest proportion of observations in all three subgenera, accounting for 46.8% in *Membranacea*, 50.0% in *Tethyos*, and 52.2% in *Vanilla*. Across the complete dataset, 319 of the 629 observations (50.7%) occurred in quadrant IV, compared with 19.4% in quadrant I, 13.0% in quadrant II, and 16.9% in quadrant III. Consistent with this distribution, mean G3/(G3+C3) values were slightly above 0.5 in all three subgenera (0.545, 0.544, and 0.544 for *Membranacea*, *Tethyos*, and *Vanilla*, respectively), whereas mean A3/(A3+T3) values remained below 0.5 (0.473, 0.476, and 0.474, respectively). Overall, the three subgenera displayed comparable patterns of nucleotide asymmetry at the third codon position.

Selective pressure analysis

Selective pressures were evaluated for 37 protein-coding genes across the 17 *Vanilla* plastomes using nonsynonymous (Ka) and synonymous (Ks) substitution rates and their ratio (Ka/Ks, ω) (Figure 9; Table S9). Based on gene-wise mean Ka/Ks values, 24 of the 37 genes (64.86%) showed ratios below 0.5, indicating strong purifying selection, whereas seven genes (18.92%) had values between 0.5 and 1.0. The remaining six genes (16.22%) exhibited mean Ka/Ks ratios greater than 1.0. Among the most strongly constrained genes, *psbA* and *psbD* showed particularly low mean Ka/Ks values of 0.026 and 0.043, respectively. Genes associated with photosystems I and II, the cytochrome b/f complex, Rubisco, DNA-dependent RNA polymerase, and *matK* generally displayed mean Ka/Ks values below 0.5. In contrast, *cemA*, *clpP*, *rpl2*, *rps12*, *rps14*, and *ycf2* showed mean Ka/Ks ratios above 1.0, ranging from 1.037 in *cemA* to 2.072 in *ycf2* (Figure 9D; Table S9).

The distribution of Ka and Ks estimates further showed that most gene comparisons occurred below the Ka = Ks reference line (Figure 9C), consistent with the predominance of Ka/Ks ratios below 1.0 across the dataset. Variation in selective pressure differed among functional groups and individual genes, with several ribosomal protein genes and conserved open reading frames displaying broader Ka/Ks distributions than genes involved in the core photosynthetic machinery (Figure 9A–D).

A marked difference was observed when the two accessions assigned to *V. pompona* were compared directly. Estimable Ka/Ks ratios were obtained for 23 of the 37 genes, including four genes with Ka/Ks > 1: *rpl16* (2.266), *ycf2* (1.932), *rpoB* (1.559), and *clpP* (1.045) (Table S9). In contrast, no estimable Ka/Ks ratios were obtained from the direct comparison between the two *V. planifolia* accessions. Thus, the two conspecific comparisons differed markedly in the extent of sequence divergence detected across the shared protein-coding genes.

Codon-based likelihood analyses were subsequently performed using the M1a versus M2a and M7 versus M8 site-model comparisons (Table S10). Significant likelihood-ratio tests were obtained for six genes under both model comparisons: *atpA*, *atpB*, *cemA*, *clpP*, *rpl2*, and *ycf2*. Under the M7 versus M8 comparison, the corresponding p-values were 0.0015 for *atpA*, 0.0044 for *atpB*, 0.0412 for *cemA*, 1.17×10^{-11} for *clpP*, 0.0384 for *rpl2*, and 6.47×10^{-62} for *ycf2*. Under the M8 model, Bayes Empirical Bayes (BEB) analysis identified sites with posterior probabilities >0.95 in eight genes (Table 4). The largest numbers of high-probability sites were detected in *ycf2* (47 sites) and *clpP* (12 sites), followed by *atpA* (two sites), whereas *atpB*, *cemA*, *rps12*, *rps14*, and *rps2* each contained one site. Importantly, the M7 versus M8 likelihood-ratio test provided statistically significant support for the models containing positively selected sites in *atpA*, *atpB*, *cemA*, *clpP*, and *ycf2*. Although BEB identified individual high-probability sites in *rps12*, *rps14*, and *rps2*, their corresponding M7 versus M8 likelihood-ratio tests were not significant (p = 0.061, 0.060, and 0.116, respectively). Conversely, *rpl2* showed a significant likelihood-ratio test but no individual site exceeded the BEB posterior-probability threshold used in Table 4.

Table 4. Positively selected genes and sites detected in the chloroplast genomes of the *Vanilla* genus

Gene	Model	LnL	LRT	LRT *p*-value	Positive selection sites (BEB)	Number of sites
atpA	M2a vs. M1a	-2606.15	11.53	3.14e-03	Dismiss	0
	M8 vs. M7	-2606.81	13.01	1.50e-03	460H*, 467K*	2
atpB	M2a vs. M1a	-2448.72	7.78	2.04e-02	Dismiss	0
	M8 vs. M7	-2450.27	10.87	4.37e-03	384E*	1
cemA	M2a vs. M1a	-1463.85	6.36	4.15e-02	13C*	1
	M8 vs. M7	-1463.86	6.38	4.12e-02	13C*	1
clpP	M2a vs. M1a	-2055.40	47.72	4.35e-11	24Y**, 43V*, 138S**, 141E**, 152N*, 197L*, 206R*, 210E*, 216I*	9
	M8 vs. M7	-2056.73	50.35	1.17e-11	24Y**, 43V**, 138S**, 141E**, 152N*, 163K*, 171C*, 197L**, 202E*, 206R**, 210E*, 216I*	12
rps12	M2a vs. M1a	-632.56	5.35	6.88e-02	Dismiss	0
	M8 vs. M7	-632.88	5.58	6.14e-02	19T*	1
rps14	M2a vs. M1a	-552.52	5.60	6.07e-02	Dismiss	0
	M8 vs. M7	-552.56	5.64	5.96e-02	41W*	1
rps2	M2a vs. M1a	-1437.13	4.33	1.15e-01	2G*	1
	M8 vs. M7	-1437.15	4.31	1.16e-01	2G*	1
ycf2	M2a vs. M1a	-13649.95	237.10	3.26e-52	9R*, 226L*, 574K*, 575M*, 1407T*, 1451S*, 1454Q*, 1604N**, 1607W*, 1610D**, 1611L**, 1612L*, 1613D**, 1615D**, 1616D*, 1617S*, 1619K*,1620M**, 1624S*, 1628I**, 1636L*, 1637K*, 1659Q**, 1661M*, 1662A*, 1666L*, 1690D*, 1696F**, 1820T**, 1986S**, 1987W**, 1988L*, 1995E**, 1996L**, 2004V**, 2034L*, 2050N**, 2051I**, 2052Y*, 2055D*, 2056K*, 2058D*, 2063K*, 2090T**, 2134P*, 2135S*	46
	M8 vs. M7	-13672.29	281.79	6.47e-62	9R*, 226L*, 574K*, 575M*, 1407T*, 1451S*, 1454Q*, 1604N**, 1607W*, 1610D**, 1611L**, 1612L*, 1613D**, 1615D**, 1616D*, 1617S*, 1619K*, 1620M**, 1621D*, 1624S*, 1628I**, 1636L**, 1637K*, 1659Q**, 1661M*, 1662A*, 1666L*, 1690D*, 1696F**, 1820T**, 1986S**, 1987W**, 1988L**, 1995E**, 1996L**, 2004V**, 2034L*, 2050N**, 2051I**, 2052Y*, 2055D*, 2056K*, 2058D*, 2063K*, 2090T**, 2134P*, 2135S*	47

LnL: log-likelihood value; LRT: likelihood ratio test; BEB: Bayes empirical Bayes analysis; * P>95%; ** P>99%. In positive sites, integers represent the position of the site, and letters represent the type of amino acid.

Phylogenetic analyses

Phylogenetic reconstruction based on concatenated chloroplast protein-coding sequences from 24 Orchidaceae plastomes recovered a well-resolved topology for the sampled Vanilloideae taxa (Figure 10). Maximum Likelihood and Bayesian analyses yielded congruent topologies, with most internal nodes receiving maximal or high support. Within Vanilloideae, the sampled representatives of tribes Pogonieae and Vanilleae

formed distinct lineages. The 17 *Vanilla* plastomes were recovered as a monophyletic group, clearly separated from the other sampled representatives of Vanilleae.

Within *Vanilla*, three major lineages corresponding to the sampled subgenera *Membranacea*, *Tethyos*, and *Vanilla* were recovered. Subgenus *Membranacea*, represented by *V. mexicana*, *V. acuta*, and *V. armoriquensis*, formed a distinct clade, with the branch subtending this lineage receiving support values of 83/0.97/0.99 for Felsenstein bootstrap, Transfer Bootstrap Expectation, and Bayesian posterior probability, respectively (Figure 10). Subgenus *Tethyos* was resolved into two sectional lineages. Section Miguelia comprised *V. shenzhenica*, *V. aphylla*, and *V. somae*, whereas section Aphyllae was represented by *V. madagascariensis*. These lineages were recovered as closely related within a distinct *Tethyos* clade. Subgenus *Vanilla* contained the largest number of sampled taxa and showed greater internal structuring. The two *V. planifolia* accessions clustered closely together, with the corresponding node receiving support values of 97/0.97/0.99. *Vanilla × tahitensis* was placed immediately adjacent to the *V. planifolia* lineage. The previously published *V. pompona* plastome (NC_036809.1) was also recovered within this upper portion of the subgenus *Vanilla* clade, close to *V. planifolia* and *V. × tahitensis*. In contrast, the *V. pompona* plastome generated in the present study (PX058849.1) occupied a clearly different phylogenetic position, grouping with the lineage containing *V. calyculata* and *V. chamissonis*. Thus, the two plastomes currently assigned to *V. pompona* were not recovered as a monophyletic pair and occupied distinct positions within subgenus *Vanilla* (Figure 10).

Overall, the plastid CDS phylogeny recovered the three sampled infrageneric lineages of *Vanilla* and further resolved sectional structure within subgenus *Tethyos*. The most conspicuous intraspecific discordance involved the two accessions assigned to *V. pompona*, whereas the two *V. planifolia* accessions showed close phylogenetic correspondence.

Discussion

The comparative analysis of 17 *Vanilla* plastomes reveals a genome that is broadly conserved in its basic architecture but considerably more dynamic at finer structural and evolutionary scales than might be expected from its overall organization. All plastomes retained the typical quadripartite configuration, whereas variation in genome size, IR/SC boundaries, gene content, repetitive sequences, sequence divergence, and large-scale rearrangements differentiated the major infrageneric lineages. Particularly notable were the enlarged plastomes and the approximately 39-kb LSC inversion shared by species of subgenus *Membranacea*, the extensive reduction of plastid *ndh* genes across the genus, and the phylogenetic discordance between the two plastomes currently assigned to *V. pompona*. Together, these patterns indicate that plastome evolution in *Vanilla* has involved both strong conservation of core chloroplast functions and lineage-specific structural change.

Plastome architecture and gene-content evolution in *Vanilla*

Although chloroplast genomes are generally more structurally conserved than their nuclear counterparts, considerable variation in genome organization has been documented across Orchidaceae (Kim et al. 2020, 2023; Guo et al. 2021). The *Vanilla* plastomes examined here conform to this general pattern. Overall GC content was remarkably stable among species, whereas genome size differed substantially among subgenera, with *Membranacea* harboring the largest plastomes. Previous studies of Vanilloideae have similarly reported variation in LSC, SSC, and IR dimensions, particularly among species with contrasting plastome architectures (Kim et al. 2023; Zhou et al. 2023; Lee et al. 2025). The higher GC content observed in the IR regions is also consistent with the widespread enrichment of GC-rich rRNA genes within these repeats and with the comparatively low evolutionary rates generally associated with IR-localized sequences (Ravi et al. 2008; Li et al. 2016a).

A particularly conspicuous feature of the *Vanilla* plastomes was the extensive reduction of the plastid NADH dehydrogenase gene complement. All canonical plastid *ndh* genes were absent except *ndhB*, which occurred as a pseudogene in all analyzed taxa. Loss and pseudogenization of plastid *ndh* genes are recurrent phenomena in Orchidaceae and have been reported in several phylogenetically distant orchid lineages, including representatives of Vanilloideae (Kim et al. 2015, 2020, 2023; Kim and Chase 2017; Zhou et al. 2023; Lee et al. 2025; Ringelmann and Wagner 2025). Because the plastid NDH complex participates in cyclic electron transport around photosystem I and contributes to photoprotection rather than being indispensable for basic photosynthesis, loss of individual plastid *ndh* genes can apparently be tolerated in several plant lineages (Sabater 2021; Petersen 2026).

Several mechanisms have been proposed to account for the recurrent loss of *ndh* genes in orchids, including altered trophic strategies, relaxed functional constraints, intracellular gene transfer, and structural instability around IR/SC boundaries (Kim et al. 2015; Lin et al. 2017; Zhou et al. 2023). However, the widespread loss observed here across photosynthetic *Vanilla* lineages suggests that it should not be interpreted exclusively as a consequence of heterotrophy. Rather, the pattern appears to represent a broader evolutionary feature of orchid plastomes whose functional consequences may differ among lineages. Moreover, transfer to the nuclear or mitochondrial genome cannot be inferred from plastome absence alone; previous orchid studies have shown that plastid *ndh* loss is not necessarily accompanied by the retention of functional transferred copies elsewhere in the cell (Lin et al. 2015, 2017).

Additional pseudogenization in *Tethyos*, involving *ccsA*, *accD*, *ycf1*, *psbC*, and *rps7* in different species, further illustrates the heterogeneous evolutionary trajectories of plastid genes within *Vanilla*. Although pseudogenes have increasingly been recognized as potentially relevant sources

of regulatory RNAs in plants (Xie et al. 2019; Yadav et al. 2023), no functional role can currently be inferred for these *Vanilla* sequences. Their taxonomically restricted occurrence is nevertheless informative because it indicates that gene degradation has continued independently after the divergence of the major infrageneric lineages.

Repetitive sequences

The predominance of mononucleotide SSRs, particularly A- and T-rich motifs, agrees with the AT-rich composition generally reported for angiosperm and orchid plastomes (Liu et al. 2025). Their preferential accumulation in the LSC region is also consistent with the greater representation of non-coding DNA in this region and with previous reports showing that plastid SSRs are disproportionately concentrated in single-copy and intergenic regions (Kumar et al. 2015; Zheng et al. 2025). Conversely, the comparatively low number of SSRs in the SSC and IR regions likely reflects both their smaller genomic extent and the generally greater structural conservation of the IRs.

The detailed analysis of mononucleotide motifs also revealed an informative contrast between conspecific accessions. The two *V. planifolia* plastomes had identical mononucleotide profiles, the same total SSR abundance, and identical distributions among structural regions. In contrast, the two accessions currently identified as *V. pompona* differed substantially in both total SSR abundance and motif composition and did not cluster together. SSR profiles alone are not sufficient for taxonomic delimitation, particularly because repeat number can evolve rapidly and can vary among individuals of the same species. Nevertheless, the contrast is noteworthy because the SSR discordance in *V. pompona* is concordant with independent differences recovered from protein-coding sequence divergence and plastid phylogeny. Thus, rather than being treated as stand-alone taxonomic evidence, the repetitive-sequence data provide an additional layer of support for closer examination of the identity of the two *V. pompona* plastomes.

Tandem repeats showed a similarly heterogeneous distribution, with short repeat units overwhelmingly predominant and the LSC containing the greatest abundance. Replication slippage and misalignment are well-established mechanisms for the generation and expansion of repetitive sequences (Ellegren 2004; Macas et al. 2009), particularly in less constrained non-coding regions. However, direct comparisons of absolute TR numbers among studies should be made cautiously because repeat counts are sensitive to detection algorithms and threshold parameters. More importantly for the present study, the heterogeneous repeat profiles indicate that repeat accumulation is not simply proportional to plastome size and therefore cannot by itself explain the larger plastomes observed in *Membranacea*.

Structural divergence

The mVISTA comparisons reinforce the general conservation of protein-coding regions in *Vanilla*, while showing considerably greater divergence in intergenic spacers. Such a distribution is expected under stronger functional constraint on coding sequences and is consistent with comparative plastome studies across angiosperms (Guisinger et al. 2011; Dong et al. 2015). The comparatively elevated variability detected in *accD*, *ycf1*, and *ycf2* is also consistent with the recognized evolutionary lability of these genes.

Variation in IR/SC junctions further demonstrates that structural conservation in *Vanilla* is not uniform across the plastome. The greatest differences were concentrated around the SSC and adjacent IR boundaries, regions that have repeatedly been implicated in plastome size variation and genomic reorganization in Orchidaceae (Kim et al. 2015, 2020, 2023; Guo et al. 2021; Zhou et al. 2023). The distinct placement of *trnL-UAG* and differences involving *ccsA* among subgenera are therefore best interpreted as part of a broader history of IR expansion and contraction rather than as isolated gene movements.

The most striking structural result, however, was the approximately 39kb inversion shared by all three species of subgenus *Membranacea*. Large LSC inversions are well documented in plastid genomes, including Orchidaceae and other angiosperm lineages. For example, large inversions have been reported in *Cypripedium*, legumes, grasses, and Passifloraceae, in some cases with breakpoint regions located near tRNA genes or repetitive sequences (Ogihara et al. 1988; Palmer et al. 1987; Hiratsuka et al. 1989; Cauz-Santos et al. 2020; Zhang et al. 2022; Liu et al. 2025). Repeat-mediated illegitimate recombination has also been implicated in extensive plastome rearrangements in Fabaceae, whereas microhomology-mediated mechanisms have been proposed for inversion breakpoints in Astragalus (Wu et al. 2021; Charboneau et al. 2021).

In *Vanilla*, the inversion boundaries were located close to *trnS-GCU* and *trnS-GGA*. Their proximity makes tRNA-associated recombination a plausible mechanistic hypothesis, but our data do not demonstrate that these tRNAs mediated the rearrangement. Indeed, no SSRs or longer tandem-repeat motifs were detected immediately adjacent to the inferred breakpoints. This absence distinguishes the *Vanilla* inversion from cases in which conspicuous repeats flank the rearranged segment and suggests that either recombination involved short microhomologies below the resolution of the repeat analyses or a different mechanism generated the inversion. Accordingly, the breakpoint-associated tRNAs should be regarded as candidate features rather than demonstrated causal elements.

The occurrence of the same inversion across all three sampled *Membranacea* species is particularly noteworthy. Because it is shared across independent plastome assemblies, the rearrangement is more consistent with a lineage-associated structural character than with an isolated

assembly artifact. Nevertheless, breakpoint validation using long-read sequencing or targeted PCR would provide an important independent confirmation, particularly given the limitations of short-read data for resolving complex structural regions (Abdullah et al. 2025). If confirmed across broader sampling, this inversion could represent a useful synapomorphic plastome feature of subgenus *Membranacea*.

The distribution of nucleotide diversity provides a complementary perspective on plastome evolution. Non-coding regions contained the strongest and most frequent peaks of variation, consistent with observations in other Orchidaceae (Tao et al. 2024, 2025). Among coding sequences, *clpP*, *infA*, *rpl32*, and rps18 were comparatively variable. Plastid loci such as *rbcL*, *matK*, and *ycf1* are already widely used for orchid identification and phylogenetics (Li et al. 2016b; Zhitao et al. 2017; Smidt et al. 2020), but the loci identified here provide additional candidates for testing within *Vanilla*. Their utility should nonetheless be experimentally validated across broader sampling of populations and species, because high nucleotide diversity at the genus level does not necessarily guarantee optimal discriminatory power at lower taxonomic levels. The strong variability of *clpP* is particularly noteworthy because this locus has also proved informative in evolutionary and biogeographic applications in other plant systems (Roman & Houston 2020).

Codon usage

Codon usage was remarkably conserved among *Vanilla* subgenera. The strong preference for synonymous codons ending in A or U is consistent with the overall AT-rich composition of the plastomes and with the widespread A/U-ending bias reported for chloroplast genes in angiosperms (Wang et al. 2024). Similarity of the RSCU profiles across *Membranacea*, *Tethyos*, and *Vanilla* indicates that the broad codon preference landscape has remained relatively stable despite substantial structural divergence among these lineages.

The complementary ENC – GC3s, neutrality, and PR2 analyses help refine interpretation of this pattern. None of the analyzed genes had ENC ≤35, indicating that extreme codon bias is uncommon in the *Vanilla* plastomes. Nevertheless, approximately two-thirds of gene–plastome observations occurred below Wright's expected ENC curve. Under the ENC framework, such departures indicate that GC3s composition alone does not fully account for the observed level of codon bias (Wright 1990). This conclusion is reinforced by the neutrality plots: GC3 explained less than 10% of the variation in GC12 within each subgenus, and regression slopes remained substantially below 1. Under the framework proposed by Sueoka (1988), these shallow relationships are inconsistent with a model in which genome-wide mutational pressure alone determines nucleotide composition across codon positions.

PR2 analysis provided an additional indication of compositional disequilibrium. Approximately half of the observations occurred in quadrant IV, corresponding to greater use of G than C and T than A at synonymous third positions, while the mean values also deviated from the neutral center of 0.5. Departure from PR2 symmetry is generally interpreted as evidence that complementary bases are not equally affected by the processes that shape synonymous codon use (Sueoka 1995; Parvathy et al. 2022).

Taken together, these analyses support a model in which codon usage in *Vanilla* reflects the combined influence of underlying AT-rich nucleotide composition and additional evolutionary or translational constraints. Importantly, however, ENC, neutrality, and PR2 plots are indirect descriptive approaches and do not themselves quantify the contribution of natural selection relative to mutation. Therefore, the present data do not justify attributing codon usage primarily to natural selection alone. A more conservative interpretation is that mutational/compositional bias is insufficient to explain the entire pattern and that other factors, potentially including translational selection and functional constraints, also contribute. Similar departures from purely mutation-driven expectations have been reported in other orchid plastomes (Li et al. 2025a; Wu et al. 2025).

Selective pressure analysis

The Ka/Ks analyses revealed a strong overall signature of purifying selection in *Vanilla* plastid protein-coding genes. This is particularly evident for genes belonging to the core photosynthetic machinery, including *psbA* and *psbD*, which exhibited extremely low mean Ka/Ks values. Strong conservation of photosynthetic genes is expected because amino acid changes in proteins participating directly in photosynthetic electron transport can have substantial functional consequences. Similar patterns of pervasive purifying selection have been reported in plastomes throughout Orchidaceae.

Against this strongly constrained background, a smaller set of genes showed elevated evolutionary rates. Mean Ka/Ks ratios greater than one in *cemA*, *clpP*, *rpl2*, *rps12*, *rps14*, and *ycf2* identify these genes as candidates for accelerated nonsynonymous evolution, but gene-wide averages alone should not be equated with definitive positive selection. This distinction is especially important when synonymous divergence is low, because Ka/Ks estimates can become unstable. The site-model analyses therefore provide a more informative second line of evidence.

Comparison of M7 and M8 yielded statistically significant LRTs together with high-probability BEB sites for *atpA*, *atpB*, *cemA*, *clpP*, and *ycf2*. These five genes consequently provide the strongest evidence of codon-specific positive selection in the present dataset. This result is consistent with previous reports of positive selection involving plastid genes in *Paphiopedilum*, *Coelogyninae*, and *Bulbophyllum*, although the specific genes differ

among orchid lineages (Tang et al. 2021; Liu et al. 2022; Li et al. 2024). Such heterogeneity is expected if different plastid functions have experienced lineage- or environment-dependent selective regimes.

Several of these candidate genes are biologically plausible targets of altered functional constraints. Genes *atpA* and *atpB* encode core components of the chloroplast ATP synthase complex involved in photophosphorylation (Neupane et al. 2019); *clpP* encodes the plastid catalytic subunit of the Clp protease, a protein-quality-control complex important for development, cellular viability, and removal of damaged or misfolded proteins (Clarke 1999; Kuroda and Maliga 2003); and *cemA* is associated with chloroplast-envelope function and carbon assimilation, with disruption of this gene affecting photosensitivity and CO_2-dependent processes (Wang et al. 2025). Finally, *ycf2*, although still incompletely characterized, has been implicated in chloroplast protein import and proteostasis (Xing et al. 2022). Its particularly strong LRT and numerous high-probability BEB sites are therefore noteworthy, especially because *ycf2* has repeatedly emerged as an evolutionarily dynamic plastid gene in Orchidaceae and other angiosperms.

These functional associations provide biologically meaningful hypotheses, but they should not be interpreted as direct evidence of adaptation to the shaded tropical environments occupied by many *Vanilla* species. Although ATP production, proteostasis, and carbon assimilation are relevant to performance under variable light and environmental stress (Ji et al. 2012; Wen et al. 2024), the site models used here identify variation in selective pressure among codons across the phylogeny rather than assigning selection to specific ecological transitions or branches. Functional assays and branch-specific evolutionary analyses would therefore be required to test whether the inferred amino-acid changes are associated with particular habitats or ecological strategies.

The remaining CodeML results further illustrate the importance of this distinction. Genes *rps12*, *rps14*, and *rps2* contained BEB sites with posterior probabilities above 0.95 under M8, but their corresponding M7–M8 LRTs were not significant. These loci should therefore be regarded as tentative candidates rather than genes with robust evidence of positive selection. Conversely, *rpl2* showed a significant LRT but lacked individual BEB sites above the adopted posterior-probability threshold. The most defensible interpretation is thus that five genes (*atpA*, *atpB*, *cemA*, *clpP*, and *ycf2*) show convergent statistical evidence, whereas the remaining loci provide partial evidence that requires further validation.

Phylogenetic analyses

The plastid CDS phylogeny strongly supports the monophyly of the sampled *Vanilla* species and recovers three major clades corresponding to subgenera *Membranacea*, *Tethyos*, and *Vanilla*. This organization is broadly congruent with the recent infrageneric classification of Karremans et al. (2025) and the broad phylogenomic study by Damian et al (2026), which recognizes four subgenera in the genus, although subgenus Gondwana was not represented in the present plastome dataset. Within *Tethyos*, the recovery of sections Miguelia and Aphyllae as distinct lineages further indicates that complete plastid coding sequences contain sufficient signal to recover meaningful lower-level structure.

The close relationship recovered between *V. × tahitensis* and *V. planifolia* deserves more cautious interpretation. Earlier molecular studies proposed *V. planifolia* as the maternal parent of *V. × tahitensis* (Lubinsky et al. 2008), whereas more recent genome-wide evidence has suggested *V. sotoarenassi* and *V. odorata* as parental taxa, with *V. sotoarenassi* contributing the maternal genome (Favre et al. 2022). Because plastids represent only the maternally inherited component of a hybrid genome in most angiosperms, the topology recovered here cannot by itself resolve the complete hybrid origin of V. × tahitensis. Moreover, the taxonomic relationship between *V. sotoarenassi* and the broadly circumscribed *V. planifolia* complex remains relevant to interpretation of this signal. Broader sampling of nuclear and plastid genomes will therefore be necessary to reconcile these alternative hypotheses.

The strongest taxonomic signal arising from the phylogenomic analysis concerns *V. pompona*. The plastome sequenced in the present study (PX058849.1) grouped with the lineage containing *V. chamissonis* and *V. calyculata*, whereas the previously published accession NC_036809.1 occupied a markedly different position close to *V. planifolia* and *V. × tahitensis*. The discrepancy is unlikely to be dismissed as minor intraspecific plastome variation, as the same two accessions also differed markedly in SSR composition and exhibited unusually high divergence across protein-coding genes. In contrast, the two independent *V. planifolia* plastomes showed highly concordant SSR profiles, little coding-sequence divergence, and close phylogenetic placement.

Recent analyses integrating nuclear, mitochondrial, and plastid evidence have likewise emphasized the need for careful reassessment of species limits and sequence identification in *Vanilla* (Karremans et al. 2025; Damián-Parizaca et al. 2025). The discordant position of NC_036809.1 relative to PX058849.1 is therefore compatible with a taxonomic inconsistency or specimen misidentification associated with the previously published plastome (Amiryousefi et al. 2017). Nevertheless, alternative biological processes such as introgression or chloroplast capture cannot be excluded solely from plastid data. Re-examination of the voucher material associated with NC_036809.1, together with nuclear genomic data and additional authenticated *V. pompona* accessions, will be necessary before its taxonomic status can be resolved definitively.

More broadly, the congruence between plastome structure and phylogenetic relationships provides an important evolutionary context for the comparative results obtained here. The shared ~39-kb inversion reinforces the distinctiveness of *Membranacea*, while the greater conservation within subgenus *Vanilla* is consistent with its relatively compact range of plastome sizes and broadly similar gene content. Complete plastomes

therefore provide considerable resolution for reconstructing infrageneric relationships in *Vanilla*, but the *V. pompona* case also demonstrates that publicly available organelle genomes should not be treated as taxonomically infallible reference sequences. Integration of authenticated vouchers, nuclear markers, and organellar genomes will be particularly important for future studies addressing species delimitation, hybridization, and conservation within this economically and evolutionarily important orchid genus.

Conclusions

Overall, this study provides a comprehensive phylogenetic framework for the genus *Vanilla* based on comparative analyses of complete cp genomes, significantly expanding the genomic resources for Vanilloideae. Our results show a largely conserved cp genome structure with subgenus-specific variation in genome size, SSC structure, IR boundary dynamics, and structural rearrangements, particularly within the subgenus *Membranacea*. The identification of hypervariable regions, consistent patterns of codon usage bias, and signatures of positive selection in key genes highlight the combined influence of natural selection and structural constraints on the evolution of the *Vanilla* cp genome. Additionally, CDS-based phylogenetic analysis resolved infrageneric and sectional relationships while uncovering potential taxonomic inconsistencies in the previously published reference cp genome. Collectively, these findings improve the understanding of the evolutionary and taxonomic history of *Vanilla* and provide a solid molecular basis for future research in systematics, conservation, genetic traceability, and crop improvement of this economically important orchid.

Declarations

Data availability

The data presented in the study are deposited in the GenBank repository, accession numbers PV739282 (*V. odorata*), PX058851 (*V. phaeantha)*, PX058849 (*V. pompona*), PX058850 (*V. chamissonis*), PX058852 (*V. calyculata*), PX659495 (*V. weberbaueriana*), PX659496 (*V. acuta*), PX659497 (*V. armoriquensis*).

Funding

This work was funded by the Peruvian Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica (CONCYTEC) and the Programa Nacional de Investigación Científica y Estudios Avanzados (PROCIENCIA), “Filogenómica de las especies peruanas del género *Vanilla* (Orchidaceae): desentrañando la historia evolutiva de la especie más popular del mundo” under grant number PE501082530-2023; by the project CUI N.° 2590699 “Mejoramiento del servicio de promoción de la ciencia, tecnología e innovación tecnológica en centro de investigación en fruticultura”, supported by the National University Toribio Rodríguez de Mendoza of Amazonas (UNTRM); by the Fundação de Apoio a Pesquisa do Distrito Federal (FAPDF), process 00193–00000229/2021-21, and Embrapa 10.22.00.014.00.00, Project “Ferramentas moleculares para caracterização da diversidade genética e identificação de espécies de baunilhas brasileiras de importância para a gastronomia e agroindústria”. This work was also supported by Vicerrectorado de Investigación de la Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas.

Acknowledgments

We extend our sincere gratitude to the Office of Information Technology (OTI) of the National University Toribio Rodríguez de Mendoza of Amazonas (UNTRM) for providing access to high-performance server infrastructure.

Author contribution statement

MZP: Conceptualization, Investigation, Methodology, Writing-original draft, Writing-review & editing. MPF: Conceptualization, Formal analysis, Writing-original draft, Writing-review & editing. ADP: Supervision, Writing – review & editing, Resources, Funding acquisition. MCRP: Funding acquisition, Investigation, Resources, Writing – review & editing. DT: Data curation, Software, Formal analysis. JRDV: Conceptualization, Supervision, Formal analysis, Writing – review & editing. ADHA: Data curation, Software, Formal analysis. CLS: Software, Formal analysis, Writing – review & editing. JJTA: Investigation, Formal analysis**,** Writing—original draft, Writing-review & editing. **JCGA** Supervision, Formal analysis, Writing – review & editing. MOC: Supervision, Resources, Funding acquisition, Project administration.

Ethics declarations

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Abdullah HL, Yan R, Chen C, Islam M, Majid A, Ahmed I, Heidari P, Tian X (2025) Evolutionary dynamics of the chloroplast genome in *Daphne* (Thymelaeaceae): Comparative analysis with related genera and insights into phylogenetics. FEBS Open Bio. https://doi.org/10.1002/2211-5463.70143
2. Álvarez-Carretero S, Kapli P, Yang Z (2023) Beginner's Guide on the Use of PAML to Detect Positive Selection. Mol Biol Evol 40(4):msad041. https://doi.org/10.1093/molbev/msad041
3. Amiryousefi A, Hyvönen J, Poczai P (2017) The plastid genome of Vanillon (*Vanilla pompona*, Orchidaceae). Mitochondrial DNA B Resour 2(2):689–691. https://doi.org/10.1080/23802359.2017.1383201
4. Amiryousefi A, Hyvönen J, Poczai P (2018) IRscope: An online program to visualize the junction sites of chloroplast genomes. Bioinformatics 34(17):3030–3031. https://doi.org/10.1093/bioinformatics/bty220
5. Beier S, Thiel T, Münch T, Scholz U, Mascher M (2017) MISA-web: A web server for microsatellite prediction. Bioinformatics 33(16):2583–2585. https://doi.org/10.1093/bioinformatics/btx198
6. Benson G (1999) Tandem repeats finder: A program to analyze DNA sequences. Nucleic Acids Res 27(2):573–580. https://doi.org/10.1093/nar/27.2.573
7. Besse P, Da Silva D, Grisoni M (2021) Plant DNA Barcoding Principles and Limits: A Case Study in the Genus Vanilla. In Besse P (ed) Molecular Plant Taxonomy: Methods and Protocols. Springer, US, pp. 131–148. https://doi.org/10.1007/978-1-0716-0997-2_8
8. Bock R, Khan MS (2004) Taming plastids for a green future. Trends Biotechnol 22(6):311–318. https://doi.org/10.1016/j.tibtech.2004.03.005
9. Bramel P, Frey F (2021) *Global strategy for the conservation and use of Vanilla genetic resources*. Zenodo. https://doi.org/10.5281/zenodo.7544770. Accessed 13 August 2026
10. Bushnell, B. (2014) BBMap: A Fast, Accurate, Splice-Aware Aligner. Lawrence Berkeley National Laboratory. https://escholarship.org/uc/item/1h3515gn. Accessed 13 August 2026
11. Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA (2012) Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics 28(4):464–469. https://doi.org/10.1093/bioinformatics/btr703
12. Cauz-Santos LA, da Costa ZP, Callot C, Cauet S, Zucchi MI, Bergès H, van den Berg C, Vieira MLC (2020) A Repertory of Rearrangements and the Loss of an Inverted Repeat Region in *Passiflora* Chloroplast Genomes. Genome Biol Evol 12(10):1841–1857. https://doi.org/10.1093/gbe/evaa155
13. Chambers A, Cibrián-Jaramillo A, Karremans AP, Moreno Martinez D, Hernandez-Hernandez J, Brym M, Resende MFR, Moloney R, Sierra SN, Hasing T, Alomia YA, Hu Y (2021) Genotyping-By-Sequencing diversity analysis of international *Vanilla* collections uncovers hidden diversity and enables plant improvement. Plant Sci 311:111019. https://doi.org/10.1016/j.plantsci.2021.111019
14. Charboneau JLM, Cronn RC, Liston A, Wojciechowski MF, Sanderson MJ (2021) Plastome Structural Evolution and Homoplastic Inversions in Neo-Astragalus (Fabaceae). Genome Biol Evol 13(10):evab215. https://doi.org/10.1093/gbe/evab215
15. Charif D, Lobry JR (2007) SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. In Bastolla U, Porto M, Roman HE, Vendruscolo M (eds), Structural Approaches to Sequence Evolution: Molecules, Networks, Populations. Springer, pp. 207–232. https://doi.org/10.1007/978-3-540-35306-5_10
16. Chen J, Wang F, Zhou C, Ahmad S, Zhou Y, Li M, Liu Z, Peng D (2023) Comparative Phylogenetic Analysis for *Aerides* (Aeridinae, Orchidaceae) Based on Six Complete Plastid Genomes. Int J Mol Sci 24(15):12473. https://doi.org/10.3390/ijms241512473
17. Chen Y, Liu N, Wang H, Luo H, Xie Z, Yu H, Chang Y, Zheng B, Zheng X, Sheng J, Jiang Y, Ye S, Hua Y, Ma H, Li F (2025) Comparative chloroplast genomics of Rutaceae: Structural divergence, adaptive evolution, and phylogenomic implications. Front Plant Sci 16:1675536. https://doi.org/10.3389/fpls.2025.1675536
18. Clarke AK (1999) ATP-dependent Clp Proteases in Photosynthetic Organisms—A Cut Above the Rest! Ann Bot 83(6):593–599. https://doi.org/10.1006/anbo.1999.0878
19. Comeron JM, Aguadé M (1998) An Evaluation of Measures of Synonymous Codon Usage Bias. J Mol Evol 47(3):268–274. https://doi.org/10.1007/PL00006384
20. Damian A, Mitidieri N (2020) Living in the clouds: A new high-elevation species of *Vanilla* (Orchidaceae, Vanilloideae) from Perú. Phytotaxa 451(2):154–160. https://doi.org/10.11646/phytotaxa.451.2.5
21. Damián-Parizaca A, Karremans AP, Barona-Colmenares AA, Iturralde G, Pérez ÁJ, de la Cruz W, Jost L, Garzón HX, Villalba-Vargas D, Zenteno-Ruiz FS, Moya-Huanca AL, Castro C, Pizango J, Vilcatoma E, Bentley AG, Jiménez MM, Baquero L, Mitidieri-Rivera N, Cameron KM (2025) The Genus *Vanilla* (Orchidaceae) in South America—Bolivia, Colombia, Ecuador, and Peru: An Annotated Checklist with a New Species. Ann Missouri Bot Gard 110(1):335–362. https://doi.org/10.3417/2025992
22. Damián-Parizaca A, Karremans AP, Pupulin F (2025) Overlooked for almost 200 years: Clarification of *Vanilla guianensis* (Orchidaceae), a prior name for *V. hostmannii*, and the reinstatement of *V. acuta*. Lankesteriana 25(3):239–251. https://doi.org/10.15517/xw71n393

23. Damián-Parizaca A, Pérez-Escobar O, Karremans AP, Antonelli A, Janovec JP, Mitidieri-Rivera N, Fitzpatrick OJ, Barona A, Wu X, Engels ME, Miranda MR, de la Cruz W, Carnevali G, Salazar G, Hágsater E, Pappas MCR, Coayla D, Tamayo-Cen I, Menchaca R, Smidt E, Lozano-Rodriguez MA, Ruiz Y, Vélez L, Garzón H, Baquero L, Iturralde G, Pérez AJ, Jiménez M, Oliva S, Cameron K (2026) Dissetangling the Vine: Phylogenomics and Historical Biogeography of Vanilla (Orchidaceae). bioRxiv. https://doi.org/10.64898/2026.04.07.716943
24. Darling ACE, Mau B, Blattner FR, Perna NT (2004) Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. Genome Res 14(7):1394–1403. https://doi.org/10.1101/gr.2289704
25. Diamond A, Barnabé S, Desgagné-Penix I (2023) Is a spice missing from the recipe? The intracellular localization of vanillin biosynthesis needs further investigations. Plant Biol 25(1):3–7. https://doi.org/10.1111/plb.13465
26. Dierckxsens N, Mardulyn P, Smits G (2017) NOVOPlasty: De novo assembly of organelle genomes from whole genome data. Nucleic Acids Res 45(4):e18. https://doi.org/10.1093/nar/gkw955
27. Dong W, Liu J, Yu J, Wang L, Zhou S (2012) Highly Variable Chloroplast Markers for Evaluating Plant Phylogeny at Low Taxonomic Levels and for DNA Barcoding. PLoS One *7*(4):e35071. https://doi.org/10.1371/journal.pone.0035071
28. Dong W, Xu C, Li C, Sun J, Zuo Y, Shi S, Cheng T, Guo J, Zhou S (2015) *ycf1*, the most promising plastid DNA barcode of land plants. Sci Rep 5(1):8348. https://doi.org/10.1038/srep08348
29. Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull. https://worldveg.tind.io/record/33886
30. Ellegren H (2004) Microsatellites: Simple sequences with complex evolution. Nat Rev Genet 5(6):435–445. https://doi.org/10.1038/nrg1348
31. Favre F, Jourda C, Grisoni M, Piet Q, Rivallan R, Dijoux J-B, Hascoat J, Lepers-Andrzejewski S, Besse P, Charron C. (2022) A genome-wide assessment of the genetic diversity, evolution and relationships with allied species of the clonally propagated crop *Vanilla planifolia* Jacks. Ex Andrews. Genet Resour Crop Evol 69(6):2125–2139. https://doi.org/10.1007/s10722-022-01362-1
32. Flanagan NS, Navia-Samboni A, Vargas WG, Díaz-Rueda DM, Cruz JYSR, Pansarin E, Castaño A (2025) Some taxonomic clarifications in *Vanilla* subgenus *Xanata* (Orchidaceae), and the resurrection of *Vanilla calyculata*. Kew Bull 80(1):83–96. https://doi.org/10.1007/s12225-024-10210-7
33. Flouri, T, Izquierdo-Carrasco F, Darriba D, Aberer AJ, Nguyen L-T, Minh BQ, Von Haeseler A, Stamatakis A (2015) The Phylogenetic Likelihood Library. Syst Biol 64(2):356–362. https://doi.org/10.1093/sysbio/syu084
34. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I (2004) VISTA: Computational tools for comparative genomics. Nucleic Acids Res 32(suppl_2):W273–W279. https://doi.org/10.1093/nar/gkh458
35. Gallage, NJ, Jørgensen K, Janfelt C, Nielsen AJZ, Naake T, Duński E, Dalsten L, Grisoni M, Møller BL (2018) The Intracellular Localization of the Vanillin Biosynthetic Machinery in Pods of *Vanilla planifolia*. Plant Cell Physiol 59(2):304–318. https://doi.org/10.1093/pcp/pcx185
36. Gao F, Chen C, Arab DA, Du Z, He Y, Ho SYW (2019) EasyCodeML: A visual tool for analysis of selection using CodeML. Ecol Evol 9(7):3891–3898. https://doi.org/10.1002/ece3.5015
37. Guisinger MM, Kuehl JV, Boore JL, Jansen RK (2011) Extreme Reconfiguration of Plastid Genomes in the Angiosperm Family Geraniaceae: Rearrangements, Repeats, and Codon Usage. Mol Biol Evol 28(1):583–600. https://doi.org/10.1093/molbev/msq229
38. Guo Y-Y, Yang J-X, Bai M-Z, Zhang G-Q, Liu Z-J (2021) The chloroplast genome evolution of *Venus slipper* (*Paphiopedilum*): IR expansion, SSC contraction, and highly rearranged SSC regions. BMC Plant Biol 21(1):248. https://doi.org/10.1186/s12870-021-03053-y
39. Havkin-Frenkel D, Belanger FC (2018) Biotechnology production of vanillin. In Havkin-Frenkel D and Dudai N (eds), Handbook of Vanilla Science and Technology. John Wiley & Sons, pp 165–186. https://doi.org/10.1002/9781119377320
40. Heissl A, Betancourt AJ, Hermann P, Povysil G, Arbeithuber B, Futschik A, Ebner T, Tiemann-Boege I (2018) Length asymmetry and heterozygosity strongly influences the evolution of poly-A microsatellites at meiotic recombination hotspots. bioRxiv 431841. https://doi.org/10.1101/431841
41. Hiratsuka J, Shimada H, Whittier R, Ishibashi T, Sakamoto M, Mori M, Kondo C, Honji Y, Sun C-R, Meng B-Y, Li Y-Q, Kanno A, Nishizawa Y, Hirai A, Shinozaki K, Sugiura M (1989) The complete sequence of the rice (*Oryza sativa*) chloroplast genome: Intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. Mol Gen Genet 217(2):185–194. https://doi.org/10.1007/BF02464880
42. Hu C, Jiao Z, Deng X, Tu X, Lu A, Xie C, Jiang K, Zeng X, Liu Z-J, Huang W, Luo Y (2022) The ecological adaptation of the unparalleled plastome character evolution in slipper orchids. Front Plant Sci 13:1075098. https://doi.org/10.3389/fpls.2022.1075098
43. Hu Y, Resende MFR, Bombarely A, Brym M, Bassil E, Chambers AH (2019) Genomics-based diversity analysis of *Vanilla* species using a *Vanilla planifolia* draft genome and Genotyping-By-Sequencing. Sci Rep 9(1):3416. https://doi.org/10.1038/s41598-019-40144-1
44. Huotari T, Korpelainen H (2012) Complete chloroplast genome sequence of *Elodea canadensis* and comparative analyses with other monocot plastid genomes. Gene 508(1):96–105. https://doi.org/10.1016/j.gene.2012.07.020
45. Ji K, Wang Y, Sun W, Lou Q, Mei H, Shen S, Chen H (2012) Drought-responsive mechanisms in rice genotypes with contrasting drought tolerance during reproductive stage. J Plant Physiol 169(4):336–344. https://doi.org/10.1016/j.jplph.2011.10.010

46. Jia X, Li H, Han Y, Wang L, Lai C, Liu X, Li P, Lei Z, Zhang Y (2025) Genome-wide microsatellite characterization and their marker development and transferability in *Broussonetia* Species. BMC Genom 26(1):61. https://doi.org/10.1186/s12864-025-11238-0
47. Jin J-J, Yu W-B, Yang J-B, Song Y, dePamphilis CW, Yi T-S, Li D-Z (2020) GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol 21(1):241. https://doi.org/10.1186/s13059-020-02154-5
48. Johnson G, Canty SWJ, Lichter-Marck IH, Wagner W, Wen J (2023) Ethanol preservation and pretreatments facilitate quality DNA extractions in recalcitrant plant species. Appl Plant Sci 11(3):e11519. https://doi.org/10.1002/aps3.11519
49. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS (2017) ModelFinder: Fast model selection for accurate phylogenetic estimates. Nat. Methods 14(6):587–589. https://doi.org/10.1038/nmeth.4285
50. Karremans AP, Chinchilla IF, Rojas-Alvarado G, Cedeño-Fonseca M, Damián A, Léotard G (2020) A reappraisal of neotropical *Vanilla*. With a note on taxonomic inflation and the importance of alpha taxonomy in biological studies. Lankesteriana 20(3):395–497. https://doi.org/10.15517/lank.v20i3.45203
51. Karremans AP, Pupulin F, Damián-Parizaca A (2025) A revised infrageneric classification of Vanilla (Orchidaceae) based on phylogenetic, morphological and biogeographical evidence. Phytotaxa 715(3):207–228. https://doi.org/10.11646/phytotaxa.715.3.2
52. Katoh K, Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol 30(4):772–780. https://doi.org/10.1093/molbev/mst010
53. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28(12):1647–1649. https://doi.org/10.1093/bioinformatics/bts199
54. Kim HT, Chase MW (2017) Independent degradation in genes of the plastid ndh gene family in species of the orchid genus *Cymbidium* (Orchidaceae; Epidendroideae). PLoS One 12(11):e0187318. https://doi.org/10.1371/journal.pone.0187318
55. Kim HT, Kim JS, Moore MJ, Neubig KM, Williams NH, Whitten WM, Kim J-H (2015) Seven New Complete Plastome Sequences Reveal Rampant Independent Loss of the *ndh* Gene Family across Orchids and Associated Instability of the Inverted Repeat/Small Single-Copy Region Boundaries. PLoS One 10(11):e0142215. https://doi.org/10.1371/journal.pone.0142215
56. Kim Y-K, Cheon S-H, Hong J-R, Kim K-J (2023) Evolutionary Patterns of the Chloroplast Genome in Vanilloid Orchids (Vanilloideae, Orchidaceae). Int J Mol Sci 24(4):3808. https://doi.org/10.3390/ijms24043808
57. Kim Y-K, Jo S, Cheon, S-H, Joo M-J, Hong J-R, Kwak M, Kim, K-J (2020) Plastome Evolution and Phylogeny of Orchidaceae, With 24 New Sequences. Front Plant Sci 11:22. https://doi.org/10.3389/fpls.2020.00022
58. Ko S-S, Jhong C-M, Lin Y-J, Wei C-Y, Lee J-Y, Shih M-C (2020) Blue Light Mediates Chloroplast Avoidance and Enhances Photoprotection of *Vanilla* Orchid. Int J Mol Sci 21(21):8022. https://doi.org/10.3390/ijms21218022
59. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A (2019) RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics 35(21):4453–4455. https://doi.org/10.1093/bioinformatics/btz305
60. Kumar M, Choi JY, Kumari N, Pareek A, Kim SR (2015) Molecular breeding in Brassica for salt tolerance: Importance of microsatellite (SSR) markers for molecular breeding in *Brassica*. Front Plant Sci 6:688. https://doi.org/10.3389/fpls.2015.00688
61. Kuroda H, Maliga P (2003) The plastid *clpP1* protease gene is essential for plant development. Nature 425(6953):86–89. https://doi.org/10.1038/nature01909
62. Lee SY, Zou P, Chen Z, Suetsugu K, Liao W (2025) Plastome and transcriptome analysis reveals a degenerated photosynthesis-related pathway in the mycoheterotrophic vanilloid orchid *Cyrtosia lindleyana*. BMC Plant Biol 25(1):841. https://doi.org/10.1186/s12870-025-06850-x
63. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W (2015) MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31(10):1674–1676. https://doi.org/10.1093/bioinformatics/btv033
64. Li F-W, Kuo L-Y, Pryer KM, Rothfels CJ (2016a) Genes Translocated into the Plastid Inverted Repeat Show Decelerated Substitution Rates and Elevated GC Content. Genome Biol Evol 8(8):2452–2458. https://doi.org/10.1093/gbe/evw167
65. Li Y, Tong Y, Xing F (2016b) DNA Barcoding Evaluation and Its Taxonomic Implications in the Recently Evolved Genus *Oberonia* Lindl. (Orchidaceae) in China. Front Plant Sci 7:1791. https://doi.org/10.3389/fpls.2016.01791
66. Li L, Wang W, Zhang G, Wu K, Fang L, Li M, Liu Z, Zeng S (2023) Comparative analyses and phylogenetic relationships of thirteen *Pholidota* species (Orchidaceae) inferred from complete chloroplast genomes. BMC Plant Biol 23(1):269. https://doi.org/10.1186/s12870-023-04233-8
67. Li L, Wu Q, Zhai J, Wu K, Fang L, Li M, Zeng S, Li S (2024) Comparative chloroplast genomics of 24 species shed light on the genome evolution and phylogeny of subtribe Coelogyninae (Orchidaceae). BMC Plant Biol 24(1):31. https://doi.org/10.1186/s12870-023-04665-2
68. Li L, Yang W, Liu Q, Jiang X, Wu K, Fang L, Li M, Zeng S, Li S (2025a) Comparative analysis of 18 chloroplast genomes reveals genomic diversity and evolutionary dynamics in subtribe Malaxidinae (Orchidaceae). BMC Plant Biol 25(1):1013. https://doi.org/10.1186/s12870-025-06772-8
69. Li Q-Q, Niu Y, Zhang Z-P, Wen J, Liao C-Y (2025b) Comparative chloroplast genome analyses of Oxytropis DC. species: New insights into genome evolution and phylogenomic implications. Front Plant Sci 16:1645582 https://doi.org/10.3389/fpls.2025.1645582

70. Li T-Z, Chen, L-J, Wang M, Chen J-B, Huang J (2019) The complete chloroplast genome of *Vanilla shenzhenica* (Orchidaceae). Mitochondrial DNA B Resour 4(2):2610–2611. https://doi.org/10.1080/23802359.2019.1642165
71. Lin C-S, Chen JJW, Huang Y-T, Chan M-T, Daniell H, Chang W-J, Hsu C-T, Liao D-C, Wu F-H, Lin S-Y, Liao C-F, Deyholos MK, Wong GK-S, Albert VA, Chou M-L, Chen C-Y, Shih M-C (2015) The location and translocation of *ndh* genes of chloroplast origin in the Orchidaceae family. Sci Rep 5(1):9040. https://doi.org/10.1038/srep09040
72. Lin C-S, Chen JJW, Chiu C-C, Hsiao HCW, Yang C-J, Jin X-H, Leebens-Mack J, de Pamphilis CW, Huang Y-T, Yang L-H, Chang W-J, Kui L, Wong GK-S, Hu J-M, Wang W, Shih, M-C (2017) Concomitant loss of NDH complex-related genes within chloroplast and nuclear genomes in some orchids. Plant J 90(5):994–1006. https://doi.org/10.1111/tpj.13525
73. Liu H, Ye H, Zhang N, Ma J, Wang J, Hu G, Li M, Zhao P (2022) Comparative Analyses of Chloroplast Genomes Provide Comprehensive Insights into the Adaptive Evolution of *Paphiopedilum* (Orchidaceae). Horticulturae 8(5):391. https://doi.org/10.3390/horticulturae8050391
74. Liu H, Jacquemyn H, Wang Y, Hu Y, He X, Zhang Ying, Zhang, Yue, Huang Y, Chen W (2025) Comparative Analysis of the Chloroplast Genomes of *Cypripedium*: Assessing the Roles of SSRs and TRs in the Non-Coding Regions of LSC in Shaping Chloroplast Genome Size. Int J Mol Sci 26(8): 3691. https://doi.org/10.3390/ijms26083691
75. Liu L, Du J, Liu Z, Zuo W, Wang Z, Li J, Zeng Y (2023) Comparative and phylogenetic analyses of nine complete chloroplast genomes of Orchidaceae. Sci Rep 13(1):21403. https://doi.org/10.1038/s41598-023-48043-2
76. Li-Zhen L, Dong-Yan T, Wu-Fu D, Shu-Dong Z (2024) The chloroplast genome of Cephalanthera nanchuanica (Orchidaceae): Comparative and phylogenetic analysis with other Neottieae species. BMC Genom 25(1):1090. https://doi.org/10.1186/s12864-024-11004-8
77. Lowe TM, Chan PP (2016) tRNAscan-SE On-line: Integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res 44(W1):W54–W57. https://doi.org/10.1093/nar/gkw413
78. Lubinsky P, Cameron KM, Molina MC, Wong M, Lepers-Andrzejewski S, Gómez-Pompa A, Kim S-C (2008) Neotropical roots of a Polynesian spice: The hybrid origin of Tahitian vanilla, *Vanilla tahitensis* (Orchidaceae). Am. J. Bot., 95(8):1040–1047. https://doi.org/10.3732/ajb.0800067
79. Ma D, Tao K, Xia Y, Luo Y, Li L (2025) Complete chloroplast genome of two *Brachycorythis* (Orchidaceae) species from China: Comparative analysis and phylogenetic implications. BMC Genom 26(1):632. https://doi.org/10.1186/s12864-025-11791-8
80. Macas J, Koblížková A, Navrátilová A, Neumann P (2009) Hypervariable 3'UTR region of plant LTR-retrotransposons as a source of novel satellite repeats. Gene 448(2):198–206. https://doi.org/10.1016/j.gene.2009.06.014
81. Mahadeo K, Palama TL, Côme B, Kodja H (2021) Vanilla: Culture, Reproduction, Phytochemistry, Curing, Pest, and Diseases. In Mérillon J-M and Kodja H (eds), Orchids: Phytochemistry, biology and horticulture (pp. 329–340). Springer. https://doi.org/10.1007/978-3-030-38392-3_13
82. Morrow KH (2024) Neutral and niche theory in community ecology: A framework for comparing model realism. Biol Philos 39(1):4. https://doi.org/10.1007/s10539-024-09941-5
83. Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3(5):418–426. https://doi.org/10.1093/oxfordjournals.molbev.a040410
84. Neupane P, Bhuju S, Thapa N, Bhattarai HK (2019) ATP Synthase: Structure, Function and Inhibition. Biomol Concepts 10(1):1–10. https://doi.org/10.1515/bmc-2019-0001
85. Niu Z, Xue Q, Zhu S, Sun J, Liu W, Ding X (2017) The Complete Plastome Sequences of Four Orchid Species: Insights into the Evolution of the Orchidaceae and the Utility of Plastomic Mutational Hotspots. Front Plant Sci 8:715. https://doi.org/10.3389/fpls.2017.00715
86. Ogihara Y, Terachi T, Sasakuma T (1988) Intramolecular recombination of chloroplast genome mediated by short direct-repeat sequences in wheat species. Proc Natl Acad Sci USA 85(22):8573–8577. https://doi.org/10.1073/pnas.85.22.8573
87. Palmer JD, Osorio B, Aldrich J, Thompson WF (1987) Chloroplast DNA evolution among legumes: Loss of a large inverted repeat occurred prior to other sequence rearrangements. Curr Genet 11(4):275–286. https://doi.org/10.1007/BF00355401
88. Pansarin ER, Ferreira AWC (2022) Evolutionary disruption in the pollination system of *Vanilla* (Orchidaceae). Plant Biol 24(1):157–167. https://doi.org/10.1111/plb.13356
89. Pansarin ER, Salatino A, Pansarin LM, Sazima M (2012) Pollination systems in Pogonieae (Orchidaceae: Vanilloideae): A hypothesis of evolution among reward and rewardless flowers. Flora: Morphol. Distrib. Funct. Ecol. Plants 207(12):849–861. https://doi.org/10.1016/j.flora.2012.09.011
90. Parvathy ST, Udayasuriyan V, Bhadana V (2022) Codon usage bias. Mol Biol Rep 49(1):539–565. https://doi.org/10.1007/s11033-021-06749-4
91. Paul, P., Malakar, A. K., & Chakraborty, S. (2018). Codon usage and amino acid usage influence genes expression level. Genetica 146(1):53–63. https://doi.org/10.1007/s10709-017-9996-4
92. Petersen G (2026) Repeated loss of plastid NDH during evolution of land plants. Ann Bot 137(1):25–46. https://doi.org/10.1093/aob/mcaf198
93. Quan J, Dong-Ping Q, Zhi W, Zuo-Zhou L, Xiao-Hong Y (2021) Research progress on local adaptation in plants. Plant Sci J 39(5):559–570. https://doi.org/10.11913/PSJ.2095-0837.2021.50559

94. Ranwez V, Douzery EJP, Cambon C, Chantret N, Delsuc F (2018) MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. Mol Biol Evol 35(10):2582–2584. https://doi.org/10.1093/molbev/msy159
95. Ravi V, Khurana JP, Tyagi AK, Khurana P (2008) An update on chloroplast genomes. Pl Syst Evol 271(1):101–122. https://doi.org/10.1007/s00606-007-0608-0
96. Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet 16(6):276–277. https://doi.org/10.1016/S0168-9525(00)02024-2
97. Ringelmann VE, Wagner ND (2025) Rapid loss of plastid *ndh* genes in slipper orchids (Cypripedioideae, Orchidaceae). Front Plant Sci 16:1507415. https://doi.org/10.3389/fpls.2025.1507415
98. Roman MG, Houston R (2020) Investigation of chloroplast regions *rps16* and *clpP* for determination of *Cannabis sativa* crop type and biogeographical origin. Leg Med 47:101759. https://doi.org/10.1016/j.legalmed.2020.101759
99. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP (2012) MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. Syst Biol 61(3):539–542. https://doi.org/10.1093/sysbio/sys029
100. Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sánchez-Gracia A (2017) DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. Mol Biol Evol 34(12):3299–3302. https://doi.org/10.1093/molbev/msx248
101. Sabater B (2021) On the Edge of Dispensability, the Chloroplast *ndh* Genes. Int J Mol Sci 22(22):12505. https://doi.org/10.3390/ijms222212505
102. Sharp PM, Li W-H (1986) An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 24(1):28–38. https://doi.org/10.1007/BF02099948
103. Sharp PM, Li W-H (1987) The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15(3):1281–1295. https://doi.org/10.1093/nar/15.3.1281
104. Sheikh-Assadi M, Naderi R, Kafı M, Fatahi R, Salami SA, Shariati V (2022) Complete chloroplast genome of *Lilium ledebourii* (Baker) Boiss and its comparative analysis: Lights into selective pressure and adaptive evolution. Sci Rep, 12(1):9375. https://doi.org/10.1038/s41598-022-13449-x
105. Smidt EC, Páez MZ, Vieira LN, Viruel J, de Baura VA, Balsanelli E, de Souza EM, Chase MW (2020) Characterization of sequence variability hotspots in Cranichideae plastomes (Orchidaceae, Orchidoideae). PLoS One 15(1):e0227991. https://doi.org/10.1371/journal.pone.0227991
106. Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313. https://doi.org/10.1093/bioinformatics/btu033
107. Steenwyk JL, Iii TJB, Li Y, Shen X-X, Rokas A (2020) ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. PLoS Biol 18(12):e3001007. https://doi.org/10.1371/journal.pbio.3001007
108. Sueoka N (1988) Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci USA 85(8):2653–2657. https://doi.org/10.1073/pnas.85.8.2653
109. Sueoka N (1995) Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J Mol Evol 40(3):318–325. https://doi.org/10.1007/BF00163236
110. Tao K, Tang L, Luo Y, Li L (2025) Complete chloroplast genome of eight *Phaius* (Orchidaceae) species from China: Comparative analysis and phylogenetic relationship. BMC Plant Biol 25(1):37. https://doi.org/10.1186/s12870-024-06040-1
111. Tao K, Tao L, Huang J, Duan H, Luo Y, Li L (2024) Complete chloroplast genome structural characterization of two *Aerides* (Orchidaceae) species with a focus on phylogenetic position of *Aerides flabellata*. BMC Genom 25(1):552. https://doi.org/10.1186/s12864-024-10458-0
112. Tao L, Duan H, Tao K, Luo Y, Li Q, Li L (2023) Complete chloroplast genome structural characterization of two *Phalaenopsis* (Orchidaceae) species and comparative analysis with their alliance. BMC Genom 24(1):359. https://doi.org/10.1186/s12864-023-09448-5
113. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones, ES, Fischer A, Bock R, Greiner S (2017) GeSeq – versatile and accurate annotation of organelle genomes. Nucleic Acids Res 45(W1):W6–W11. https://doi.org/10.1093/nar/gkx391
114. Wang D, Wu Y-W, Shih AC-C, Wu C-S, Wang Y-N, Chaw S-M (2007) Transfer of Chloroplast Genomic DNA to Mitochondrial Genome Occurred At Least 300 MYA. Mol Biol Evol 24(9):2040–2048. https://doi.org/10.1093/molbev/msm133
115. Wang G, Ren Y, Su Y, Zhang H, Li J, Han J (2024) Molecular marker development and phylogenetic analysis of *Aconitum* species based on chloroplast genomes. Ind Crop Prod 221:119386. https://doi.org/10.1016/j.indcrop.2024.119386
116. Wang T, Feng H, Zhu H, Zhong B (2025) Molecular phylogeny and comparative chloroplast genome analysis of the type species *Crucigenia quadrata*. BMC Plant Biol 25(1):64. https://doi.org/10.1186/s12870-025-06070-3
117. Watteyn C, Reubens B, Bolaños JBA, Campos FS, Silva AP, Karremans AP, Muys B (2023) Cultivation potential of *Vanilla* crop wild relatives in two contrasting land use systems. Eur J Agron 149:126890. https://doi.org/10.1016/j.eja.2023.126890
118. Wen J, Zhu J-W, Ma X-D, Li H-M, Wu B-C, Zhou W, Yang J-X, Song C-F (2024) Phylogenomics and adaptive evolution of hydrophytic umbellifers (tribe Oenantheae, Apioideae) revealed from chloroplast genomes. BMC Plant Biol 24(1):1140. https://doi.org/10.1186/s12870-024-05863-2

119. Wickham H (2016) Ggplot2. Springer International Publishing. https://doi.org/10.1007/978-3-319-24277-4

120. Wolfe KH, Li WH, Sharp PM (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci USA 84(24):9054–9058. https://doi.org/10.1073/pnas.84.24.9054

121. Wong TKF, Ly-Trong N, Ren H, Demotte P, Baños H, Roger AJ, Susko E, Bielow C, Maio ND, Goldman N, Hahn MW, Reis M, Vinh LS, Huttley G, Lanfear R, Minh BQ (2026) IQ-TREE 3: Phylogenomic Inference Software using Complex Evolutionary Models. Mol Biol Evol 43(5):msag117. https://doi.org/10.1093/molbev/msag117

122. Wright F (1990) The 'effective number of codons' used in a gene. Gene 87(1):23–29. https://doi.org/10.1016/0378-1119(90)90491-9

123. Wu C, Liu R, Li C, Shang M, Zhang Y, Bai K, Wang Y, Zhou Y, Lin Y, Duan B (2025) Comparative analysis of codon usage bias and phylogenetic relationships in chloroplast genomes across 49 *Dendrobium* species. BMC Plant Biol, 25(1):1411. https://doi.org/10.1186/s12870-025-07395-9

124. Wu S, Chen J, Li Y, Liu A, Li A, Yin M, Shrestha N, Liu J, Ren G (2021) Extensive genomic rearrangements mediated by repetitive sequences in plastomes of *Medicago* and its relatives. BMC Plant Biol 21(1):421. https://doi.org/10.1186/s12870-021-03202-3

125. Wu Y, Zeng, M-Y, Wang, H-X, Lan S, Liu Z-J, Zhang S, Li M-H, Guan Y (2024) The Complete Chloroplast Genomes of *Bulbophyllum* (Orchidaceae) Species: Insight into Genome Structure Divergence and Phylogenetic Analysis. Int J Mol Sci 25(5):2665. https://doi.org/10.3390/ijms25052665

126. Wu Z, Liao R, Yang T, Dong X, Lan D, Qin R, Liu H (2020) Analysis of six chloroplast genomes provides insight into the evolution of *Chrysosplenium* (Saxifragaceae). BMC Genom 21(1):621. https://doi.org/10.1186/s12864-020-07045-4

127. Xie J, Li Y, Liu X, Zhao Y, Li B, Ingvarsson PK, Zhang D (2019) Evolutionary Origins of Pseudogenes and Their Association with Regulatory Sequences in Plants. Plant Cell 31(3):563–578. https://doi.org/10.1105/tpc.18.00601

128. Xing J, Pan J, Yi H, Lv K, Gan Q, Wang M, Ge H, Huang X, Huang F, Wang Y, Rochaix J-D, Yang W (2022) The plastid-encoded protein Orf2971 is required for protein translocation and chloroplast quality control. Plant Cell 34(9):3383–3399. https://doi.org/10.1093/plcell/koac180

129. Yadav S, Kalwan G, Meena S, Gill SS, Yadava YK, Gaikwad K, Jain PK (2023) Unravelling the due importance of pseudogenes and their resurrection in plants. Plant Physiol Biochem 203:108062. https://doi.org/10.1016/j.plaphy.2023.108062

130. Yang H, Barros-Rios J, Kourteva G, Rao X, Chen F, Shen H, Liu C, Podstolski A, Belanger F, Havkin-Frenkel D, Dixon RA (2017) A re-evaluation of the final step of vanillin biosynthesis in the orchid *Vanilla planifolia*. Phytochemistry 139:33–46. https://doi.org/10.1016/j.phytochem.2017.04.003

131. Yang Z (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. Bioinformatics, 13(5):555–556. https://doi.org/10.1093/bioinformatics/13.5.555

132. Yang Z, Nielsen R, Goldman N, Pedersen A-MK (2000) Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. Genetics 155(1):431–449. https://doi.org/10.1093/genetics/155.1.431

133. Yang Z, Wong WSW, Nielsen R (2005) Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection. Mol Biol Evol 22(4):1107–1118. https://doi.org/10.1093/molbev/msi097

134. Ye P, Wu J, An M, Chen H, Zhao X, Jin X, Si Q (2021) Geographical Distribution and Relationship with Environmental Factors of *Paphiopedilum* Subgenus Brachypetalum Hallier (Orchidaceae) Taxa in Southwest China. Diversity 13(12):634. https://doi.org/10.3390/d13120634

135. Zhang D, Gao F, Jakovlić I, Zou H, Zhang J, Li WX, Wang GT (2020) PhyloSuite: An integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. Mol Ecol Resour 20(1):348–355. https://doi.org/10.1111/1755-0998.13096

136. Zhang J-Y, Liao M, Cheng Y-H, Feng Y, Ju W-B, Deng H-N, Li X, Plenković-Moraj A, Xu B (2022) Comparative Chloroplast Genomics of Seven Endangered *Cypripedium* Species and Phylogenetic Relationships of Orchidaceae. Front Plant Sci 13:911702. https://doi.org/10.3389/fpls.2022.911702

137. Zhang Z (2022) KaKs_Calculator 3.0: Calculating Selective Pressure on Coding and Non-Coding Sequences. Genom Proteom Bioinform 20(3):536–540. https://doi.org/10.1016/j.gpb.2021.12.002

138. Zheng H-Z, Peng G-X, Zhao, L-C, Dai W, Xu, M-H, Xu X-G, Tang M (2025) Comparative and evolutionary analysis of chloroplast genomes from five rare *Styrax* species. BMC Genom 26(1):450. https://doi.org/10.1186/s12864-025-11629-3

139. Zheng S, Poczai P, Hyvönen J, Tang J, Amiryousefi A (2020) Chloroplot: An Online Program for the Versatile Plotting of Organelle Genomes. Front Genet 11:576124. https://doi.org/10.3389/fgene.2020.576124

140. Zhitao N, Shuying Z, Jiajia P, Ludan L, Jing S, Xiaoyu D (2017) Comparative analysis of *Dendrobium* plastomes and utility of plastomic mutational hotspots. Sci Rep 7(1):2073. https://doi.org/10.1038/s41598-017-02252-8

141. Zhou L, Chen T, Qiu X, Liu J, Guo S (2023) Evolutionary differences in gene loss and pseudogenization among mycoheterotrophic orchids in the tribe Vanilleae (subfamily Vanilloideae). Front Plant Sci 14:1160446. https://doi.org/10.3389/fpls.2023.1160446

142. Zhu L, Wu H, Li H, Tang H, Zhang L, Xu H, Jiao F, Wang N, Yang L (2021) Short Tandem Repeats in plants: Genomic distribution and function prediction. Electron J Biotechnol, 50:37–44. https://doi.org/10.1016/j.ejbt.2020.12.003

Figures

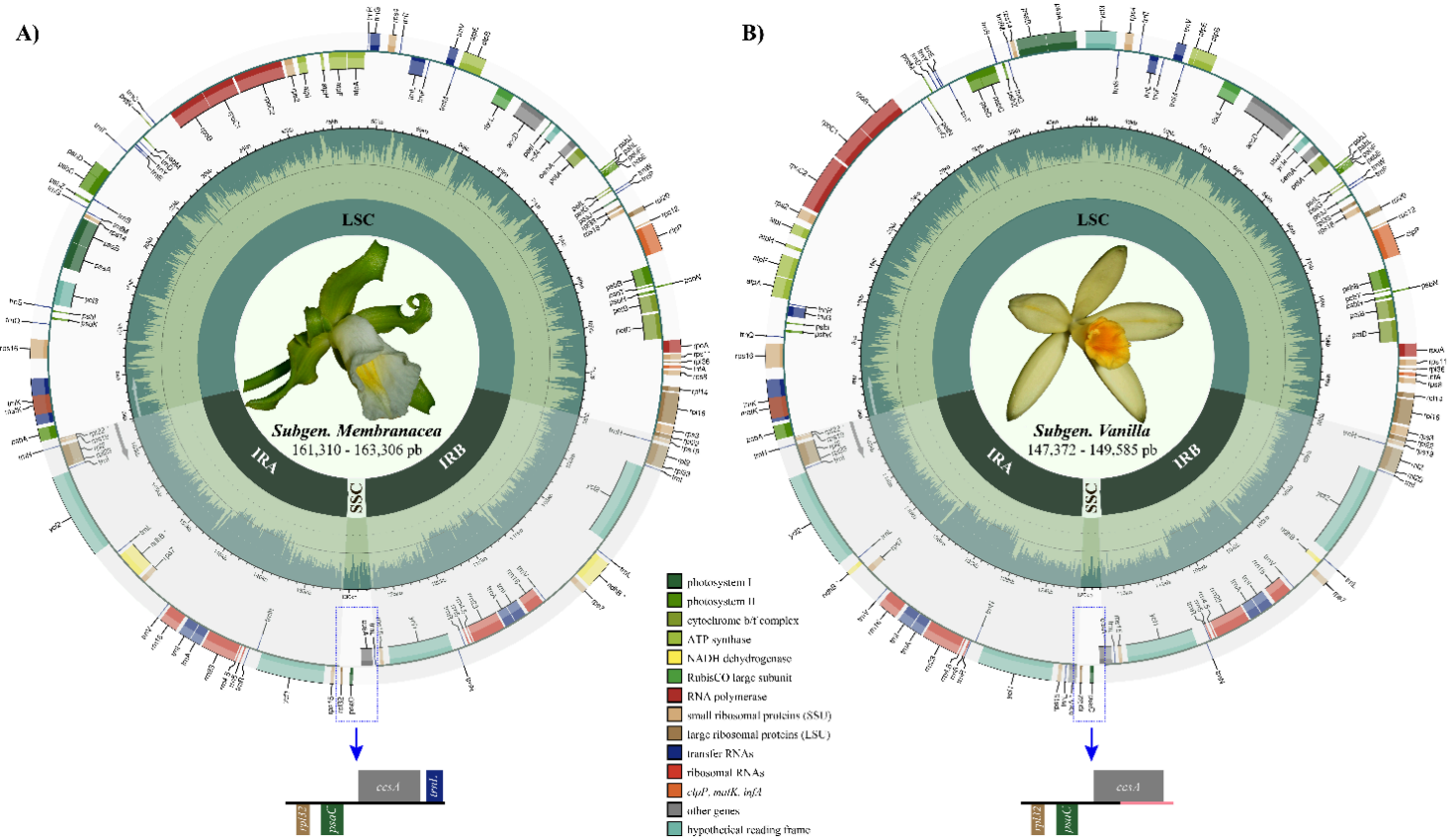

Figure 1

Circular gene maps of the chloroplast genomes of *Vanilla* species sequenced and analyzed in this study. (A) Representative cp genome of the species in the subgenus *Membranacea*. (B) Representative cp genome of the species in the subgenus *Vanilla*. Genes inside and outside the circle are transcribed clockwise and counterclockwise, respectively. Genes belonging to different functional groups are shown in different colors. Blue dashed rectangles indicate the small single-copy region in both subgenera.

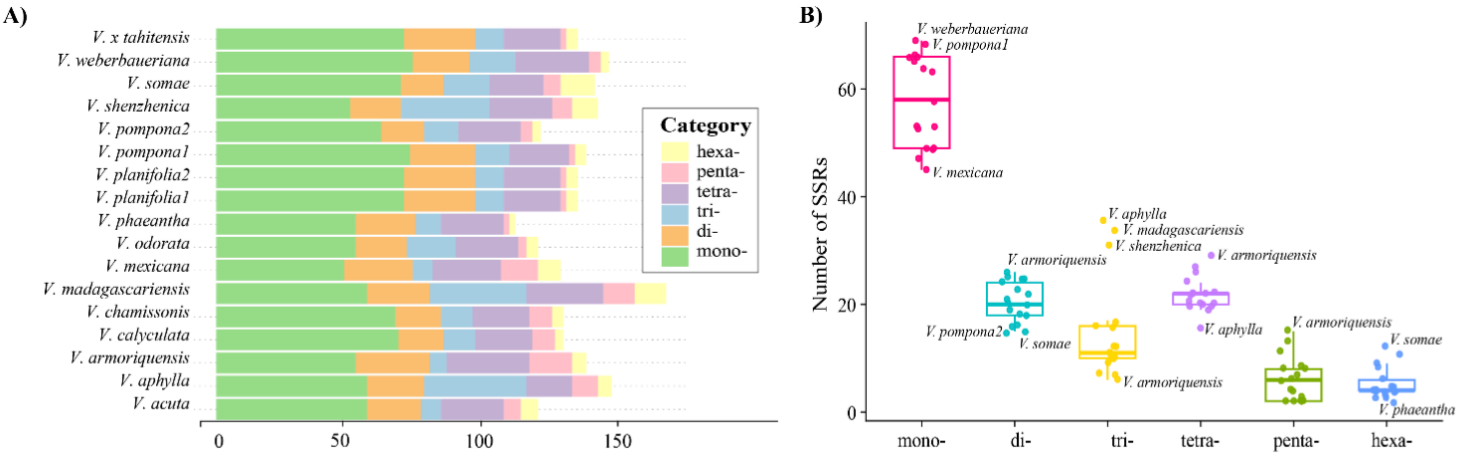

Figure 2

Distribution and composition of simple sequence repeats (SSRs) across 17 *Vanilla*chloroplast genomes. **A)** Stacked horizontal bar plot showing the contribution of mono-, di-, tri-, tetra-, penta-, and hexanucleotide SSRs to the total SSR abundance in each plastome. **B)** Box plots showing variation in the abundance of the six SSR classes across the analyzed plastomes; individual points represent chloroplast genomes. *V. planifolia1* (NC_026778); *V. planifolia2* (MN200375); *V. pompona1* (NC_036809)*; V. pompona2* (PX058849).

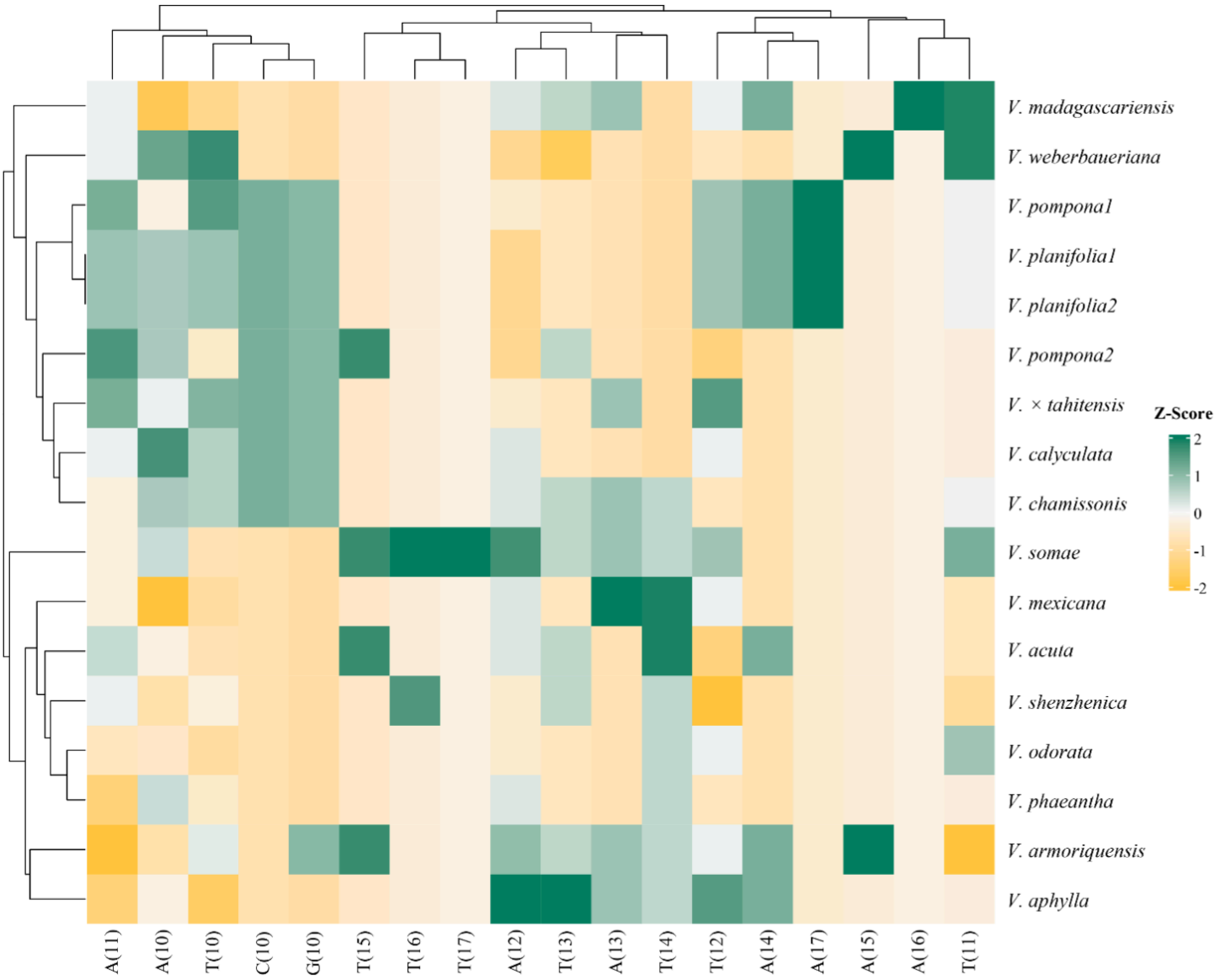

Figure 3

Hierarchical clustering of mononucleotide simple sequence repeat (SSR) profiles across 17 *Vanilla*chloroplast genomes. Rows represent individual plastomes and columns represent mononucleotide SSR motifs. Colors indicate standardized motif abundance (Z-scores), ranging from lower relative abundance (yellow) through values near the mean (white) to higher relative abundance (green). Dendrograms show hierarchical clustering of plastomes and SSR motifs based on similarities in their abundance profiles.

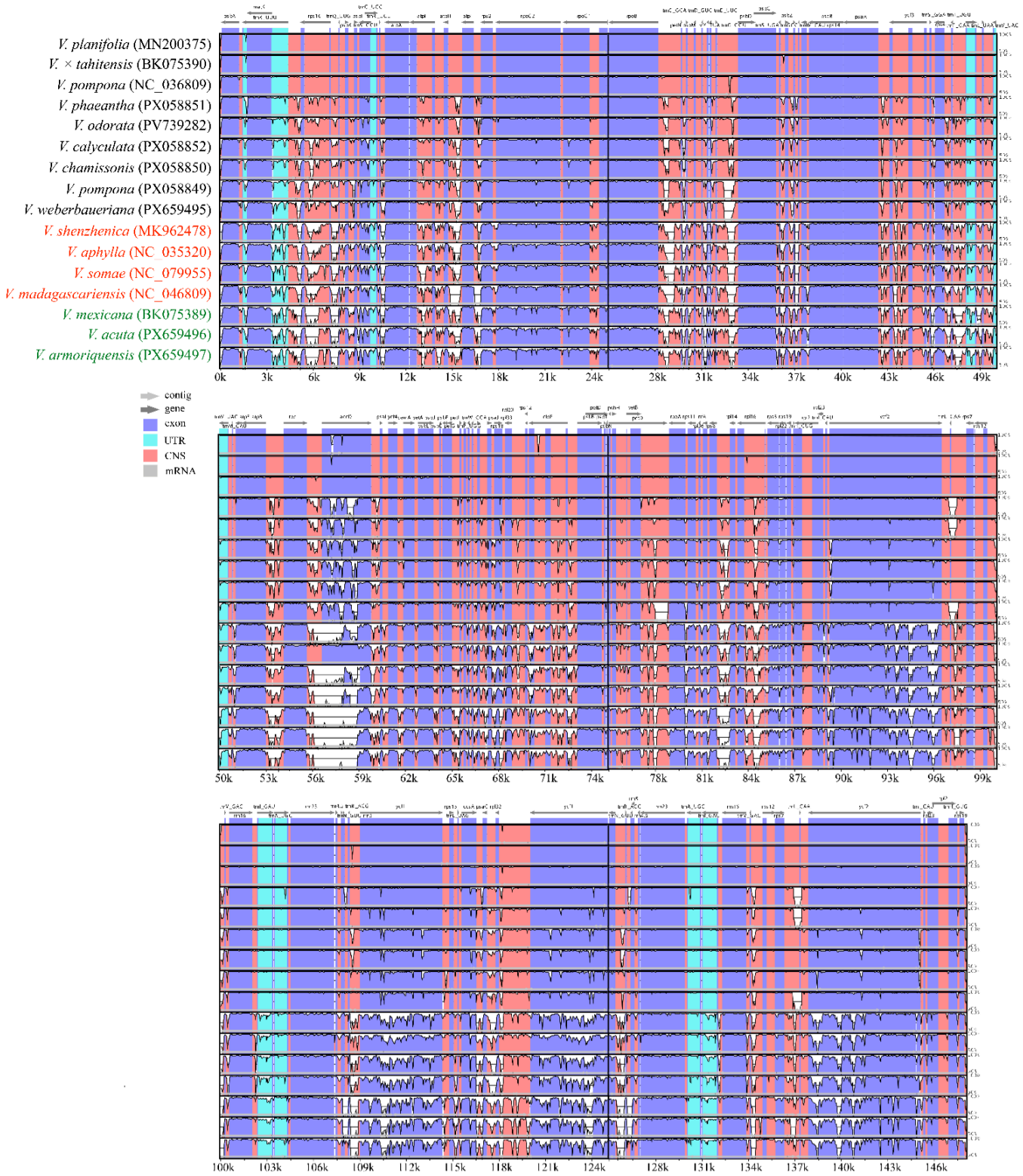

Figure 4

Global alignment of 17 *Vanilla* chloroplast genomes, with *V. planifolia*(NC_026778.1) as reference. The x-axis indicates the positions on the chloroplast genome. The y-axis shows the average percent identity of the aligned regions with the reference, ranging from 50% to 100%. Coding regions are colored purple, and non-coding regions are colored pink. Species highlighted in green correspond to the subgenus *Membranacea*, those indicated in red belong to the subgenus *Tethyos*, whereas species shown in black are assigned to the subgenus *Vanilla*.

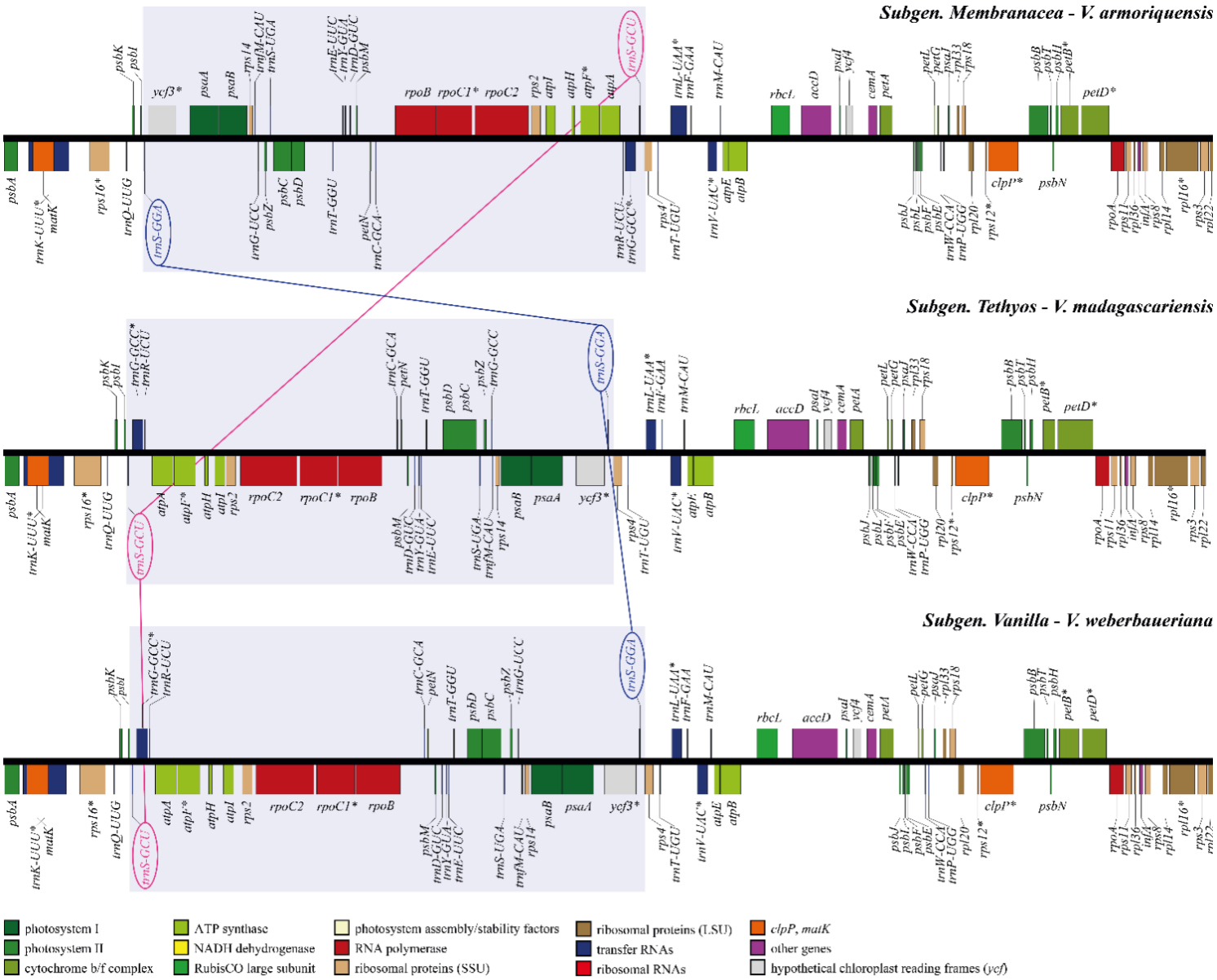

Figure 5

Schematic representation of the large single-copy (LSC) region in representative species from each *Vanilla* subgenus. The blue-shaded box highlights the inverted genomic segment and its gene arrangement. Pink and blue lines trace the positions of *trnS-GCU* and *trnS-GGA*, respectively, illustrating their involvement at the inversion breakpoints. * Genes with introns.

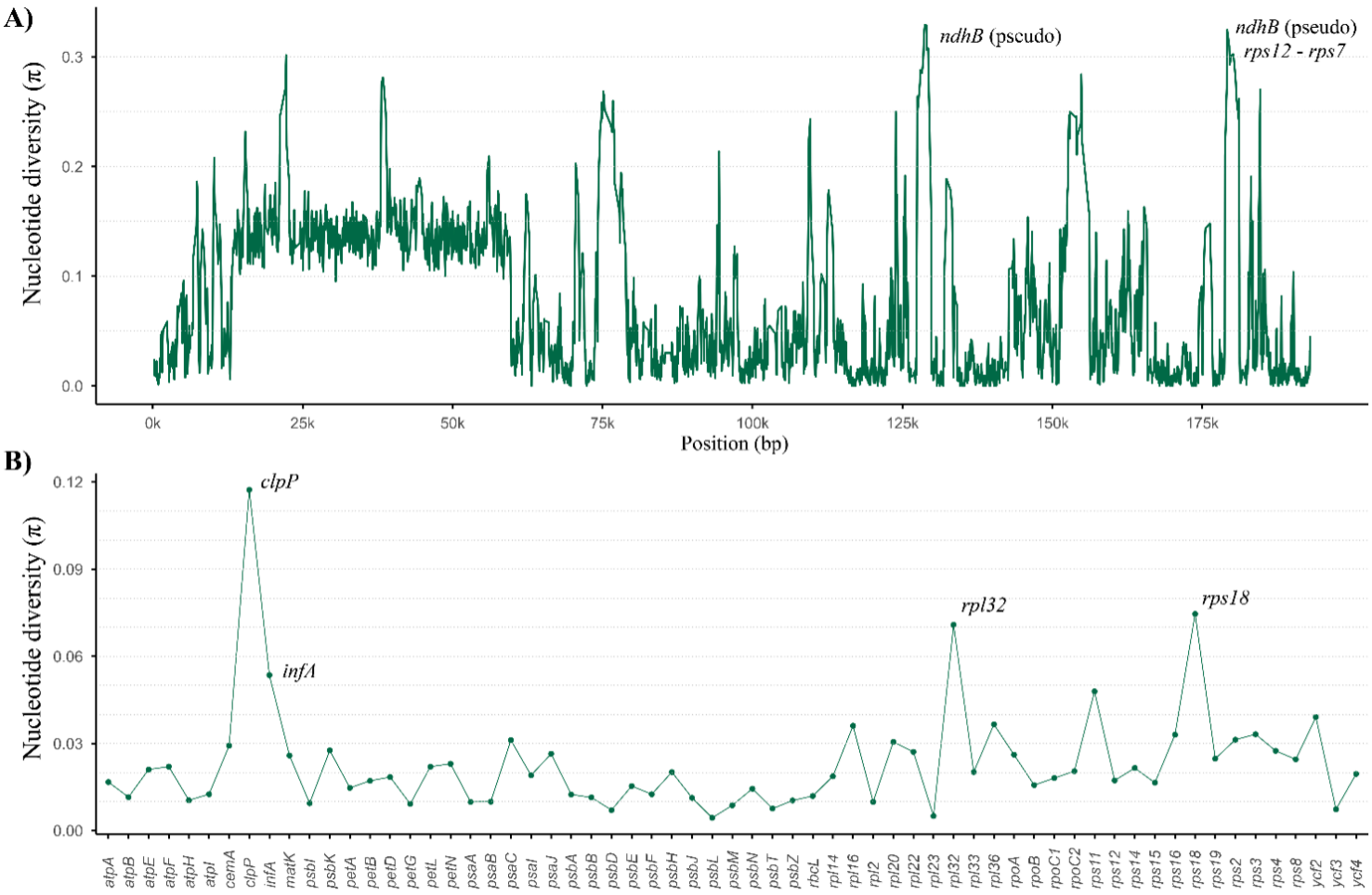

Figure 6

Nucleotide diversity across 17 *Vanilla* chloroplast genomes. **A**) Sliding-window analysis of nucleotide diversity (π) across the complete plastome alignment, highlighting regions with the highest sequence variation. **B**) Nucleotide diversity (π) of 63 shared protein-coding genes; *clpP, infA, rpl32,* and *rps18* showed π values greater than 0.05

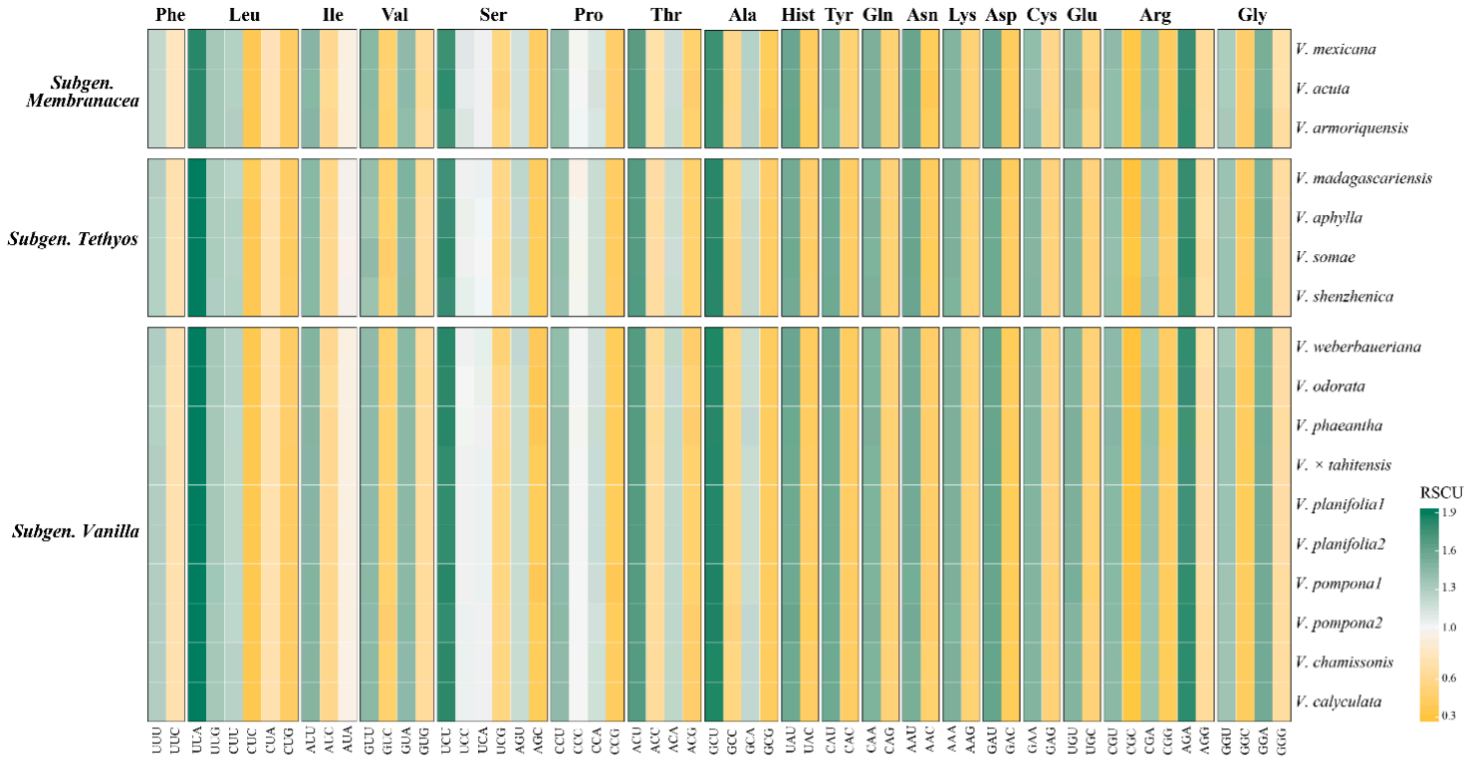

Figure 7

Heatmap of relative synonymous codon usage (RSCU) across 17 *Vanilla*chloroplast genomes. The heatmap displays 59 informative synonymous codons, excluding the three stop codons and the single-codon amino acids methionine (AUG) and tryptophan (UGG). Rows represent individual plastomes grouped by subgenus, and columns represent synonymous codons grouped by their encoded amino acid. Colors indicate RSCU values, ranging from lower codon usage (yellow) to higher codon usage (green), with values around 1 indicating no preferential usage.

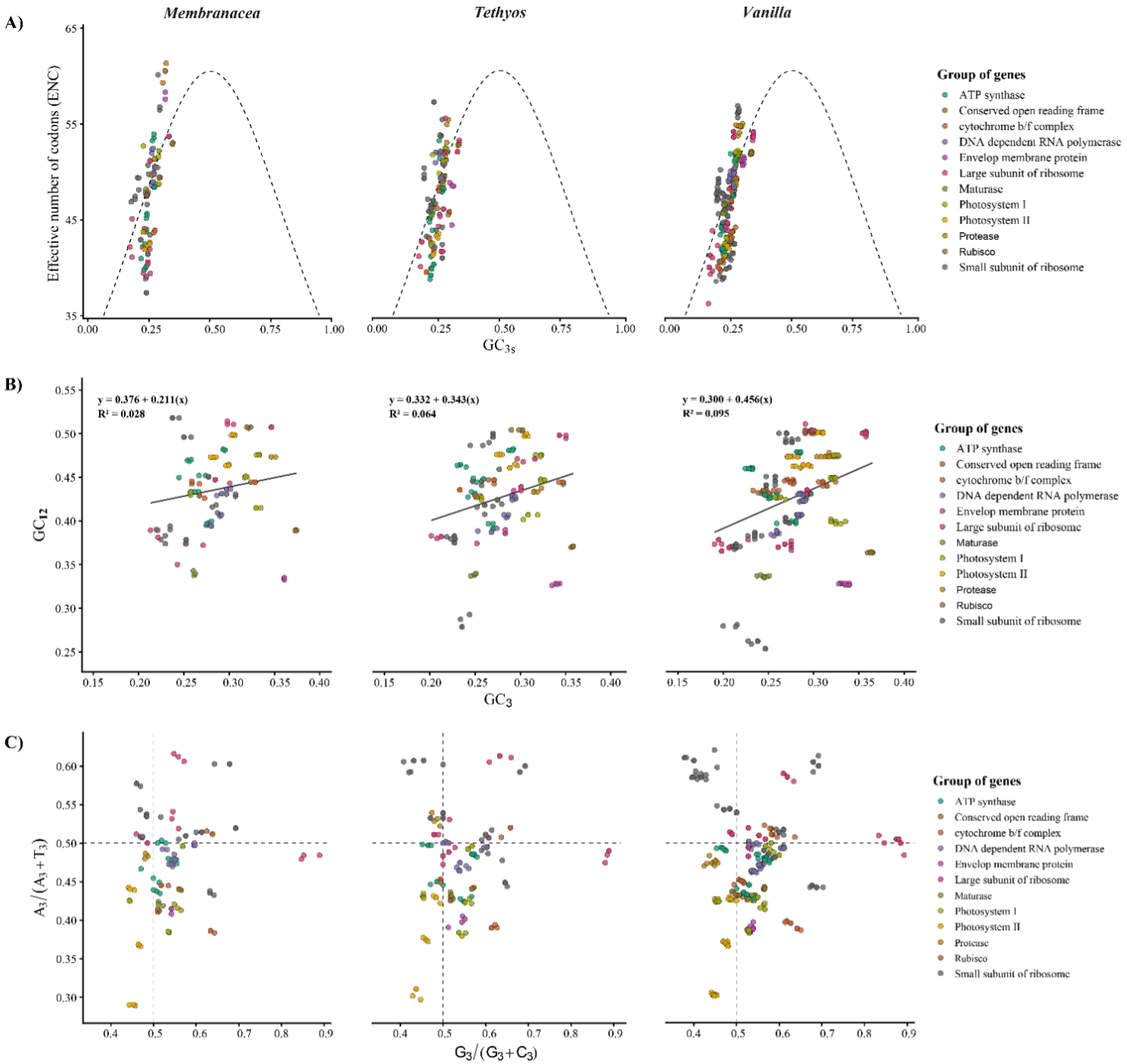

Figure 8

Comparative analysis of codon usage bias across three *Vanilla* subgenera. (A) ENC–GC3s plots showing the relationship between the effective number of codons (ENC) and GC content at synonymous third codon positions (GC3s); the dashed curve represents Wright's theoretical expectation. (B) Neutrality plots showing the relationship between GC3 and GC12, with fitted regression lines and the corresponding regression equations and coefficients of determination (R^2). (C) PR2 plots showing the relationship between G3/(G3+C3) and A3/(A3+T3); dashed lines at 0.5 indicate parity between G/C and A/T at the third codon position. Points are colored according to functional gene group, and panels are arranged by subgenus (*Membranacea*, *Tethyos*, and *Vanilla*).

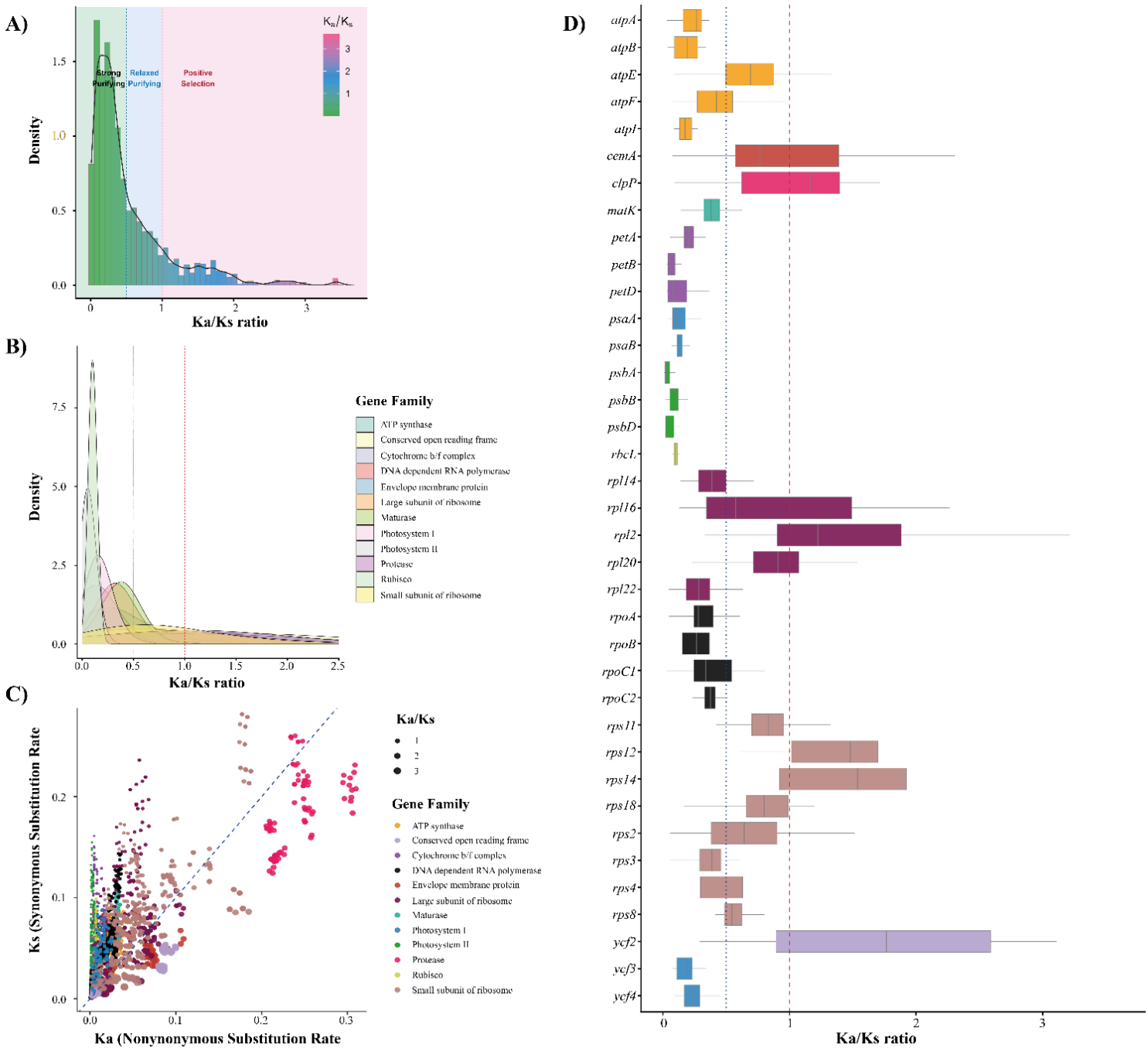

Figure 9

Comprehensive Ka/Ks analysis of chloroplast genes in *Vanilla* species. **A)** Histogram with an overlaid density curve showing the overall distribution of Ka/Ks ratios for all analyzed genes. Background shading indicates regions under different selection regimes: green for strong purifying selection (Ka/Ks ≤ 0.5), blue for relaxed purifying selection (0.5 < Ka/Ks ≤ 1.0), and pink for positive selection (Ka/Ks > 1.0). The black curve indicates the smoothed density distribution. **B)** Density plots illustrating the distribution of Ka/Ks ratios across different functional gene families. Each colored curve represents a distinct gene family, with overlapping areas highlighting differences in evolutionary constraints. Vertical dashed red and dotted blue lines indicate thresholds for neutral evolution and strong purifying selection, respectively. **C)** Scatter plot showing the relationship between synonymous (Ks) and nonsynonymous (Ka) substitution rates for individual chloroplast genes. Data points are color-coded by gene family and scaled by Ka/Ks ratio. The dashed diagonal line indicates the neutral evolution threshold (Ka = Ks). **D)** Box plots showing the distribution of Ka/Ks ratios for individual chloroplast genes, ordered by their median values. Boxes represent interquartile ranges with medians indicated by horizontal lines. The dashed red line indicates the neutral evolution threshold (Ka/Ks = 1), and the dotted blue line marks the threshold for strong purifying selection (Ka/Ks = 0.5).

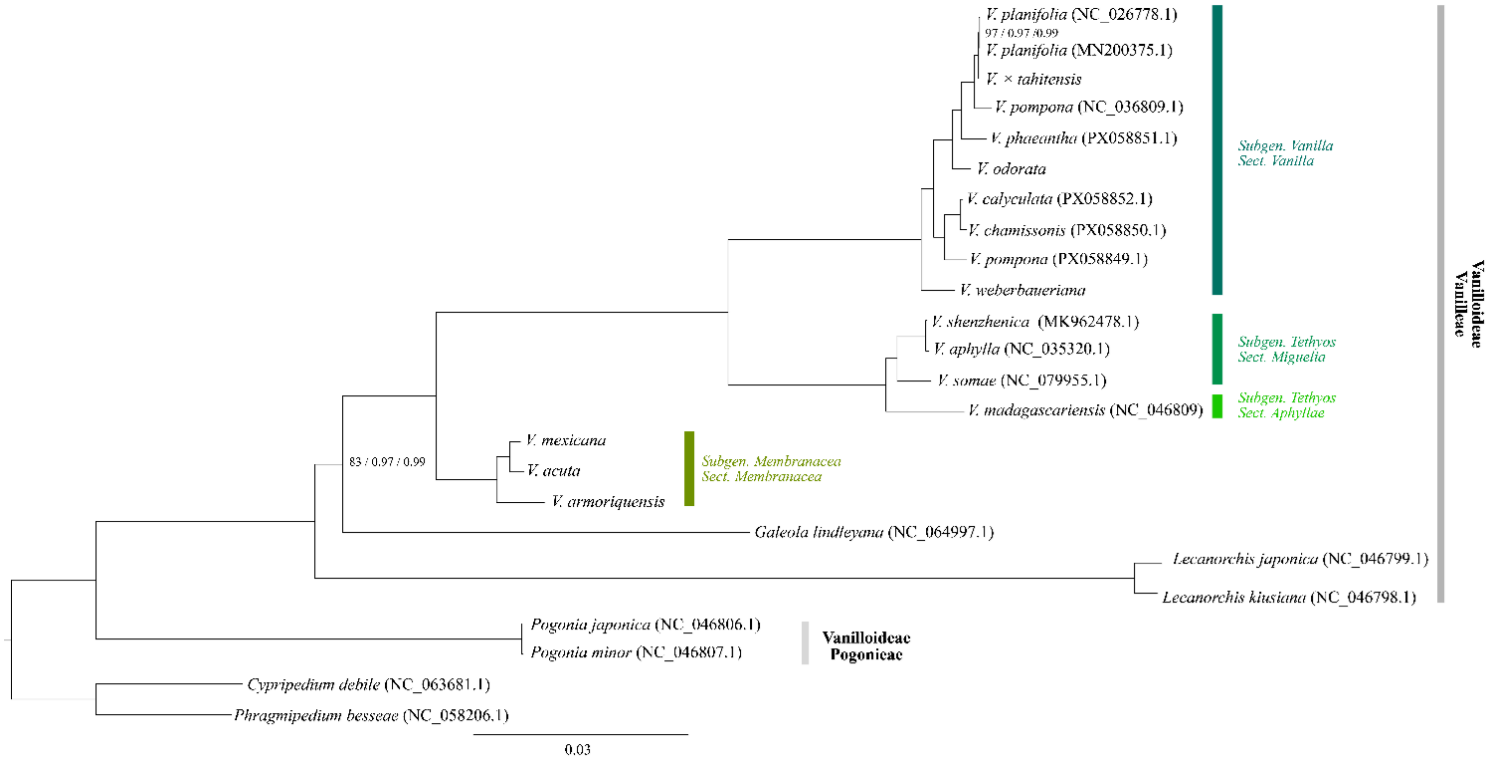

Figure 10

Maximum Likelihood phylogenetic reconstruction of *Vanilla*based on chloroplast protein-coding genes inferred using RAxML Next Generation under the GTR + F + I + R3 substitution model. Node support values are shown as follows: Felsenstein bootstrap / Transfer Bootstrap Expectation / BI posterior probability; nodes without displayed values indicate maximal support (100/1/1).

Supplementary Files

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

- ESM1.png
- ESM2.png
- SupplementaryMaterial.docx
- TableS1.xlsx
- TableS2.xlsx
- TableS3.xlsx
- TableS4.xlsx
- TableS5.xlsx
- TableS6.xlsx
- TableS7.xlsx
- TableS8.xlsx
- TableS9.xlsx
- TableS10.xlsx