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 Genome Analysis and Evolutionary Insights in the Versatile Medicinal Plant Calendula officinalis L

Ningyun Zhang
Hunan University of Arts and Science
Kerui Huang
huangkerui@huas.edu.cn
Hunan University of Arts and Science
Peng Xie
Hunan University of Arts and Science
Aihua Deng
Hunan University of Arts and Science
Ming Jiang
Hunan University of Arts and Science
Ping Mo
Hunan University of Arts and Science
Hanbin Yin
Hunan University of Arts and Science
Xuan Tang
Hunan University of Arts and Science
Rongjie Huang
Hunan University of Arts and Science
Jiale Liang
Hunan University of Arts and Science
Fuhao He
Hunan University of Arts and Science
Haoliang Hu
Hunan University of Arts and Science
Yun Wang
Hunan University of Arts and Science

Article

Keywords: Calendula officinalis, chloroplast genome, codon usage bias, evolution, adaptation

Posted Date: February 5th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3899864/v1

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

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on April 26th, 2024. See the published version at https://doi.org/10.1038/s41598-024-60455-2.

Abstract

Calendula officinalis L.is a versatile medicinal plant with numerous applications in various fields. However, its chloroplast genome structure, features, phylogeny, and patterns of evolution and mutation remain largely unexplored. This study examines the chloroplast genome, phylogeny, codon usage bias, and divergence time of *C. officinalis*, enhancing our understanding of its evolution and adaptation. The chloroplast genome of *C. officinalis* is a 150,465 bp circular molecule with a G+C content of 37.75% and comprises 131 genes. Phylogenetic analysis revealed a close relationship between *C. officinalis*, *C. arvensis*, and *Osteospermum ecklonis*. A key finding is the similarity in codon usage bias among these species, which, coupled with the divergence time analysis, supports their close phylogenetic proximity. This similarity in codon preference and divergence times underscores a parallel evolutionary adaptation journey for these species, highlighting the intricate interplay between genetic evolution and environmental adaptation in the Asteraceae family.

Introduction

Calendula officinalis L., a short-lived annual herbaceous species of the genus *Calendula* in the Asteraceae family, garners global recognition for its ubiquity and resilience. Predominantly located in the United States and Europe, it thrives in sunlit or partially shaded environments, necessitating minimal cultivation and management [1]. The plant stands 12–30 inches tall, characterized by its yellow to orange hermaphrodite flowers usually 2–3 inches in diameter, which bloom in a head-shaped inflorescence [2]. The leaves of *C. officinalis* are oblanceolate, alternate, sessile, and bright green, with stems adorned with umbel-like branches. The plant yields a curved, ring-shaped, and sickle-shaped achene [2]. The European Union has funded multiple research projects focused on *C. officinalis* due to its multifaceted role. Its diverse colors and aroma make it a favored decorative plant, and its bioactive compounds — carotenoids, saponins, amino acids — have found significant applications in chemical and pharmacological domains, offering anti-inflammatory, anti-viral, anti-genotoxic properties among others [2–3]. Despite its extensive utility, a comprehensive understanding of the chloroplast genome structure, phylogeny, and evolutionary mutation patterns of *C. officinalis* remains elusive.

Chloroplasts, critical plant organelles, govern photosynthesis, biosynthesis, and carbon sequestration [4]. These organelles possess an independent genetic system from the nuclear genome, and since the first chloroplast genome from *Nicotiana tabacum* was sequenced, the structure and function of chloroplast genomes have been progressively elucidated [5]. A typical chloroplast genome measures between 100 kb to 200 kb and has a four-part structure encompassing the large-single copy (LSC) region, the small-single copy (SSC) region, and two inverted repeat regions (IR) [6]. High-throughput sequencing technologies have facilitated the sequencing of numerous plant chloroplast genomes, thereby bolstering systematic evolutionary research based on chloroplast genome codon analysis. Codons form the fundamental link between nucleic acids and proteins, and synonymous codons, barring methionine and tryptophan, encode identical amino acids [7–8]. Codon usage bias, a phenomenon prevalent across organisms, refers to the variability in the frequency of synonymous codons coding for the same amino acid [9]. This bias not only

mirrors the species' or genes' origin, evolution, and mutation patterns but also significantly impacts gene function and protein expression [10]. Despite previous studies focusing on the codon usage bias in nuclear genomes [11–12], the chloroplast genomes' genetic code varies from the standard genetic code [13], thereby necessitating an analysis of the codon usage bias in the chloroplast genomes.

Next-generation sequencing technologies have enabled the complete sequencing of numerous chloroplast genomes, and over two thousand have been deposited in GenBank at the National Center for Biotechnology Information (NCBI). Most plant species' codon usage bias based on chloroplast genomes has been analyzed, and their phylogenetic status, evolution, and mutation patterns have been well-delineated. For instance, multiple species of the *Oryza* and *Gynostemma* genera have had their phylogenetic status, evolution, and mutation patterns elucidated through phylogenetic analysis and codon usage bias analysis based on chloroplast genomes.

In the present investigation, we characterized the chloroplast genome of *C. officinalis* and performed a comprehensive phylogenetic analysis anchored on this genome. Furthermore, we conducted a detailed exploration of the codon usage bias in *C. officinalis* and its closely related species *Calendula arvensis* and *Osteospermum ecklonis*. This approach facilitated our understanding of the plant's genomic attributes, evolutionary adaptation mechanisms, and phylogenetic positioning.

Materials and Methods

Plant materials and genome sequencing

Fresh leaves were picked from *C. officinalis* planted near the Changde Vocational Technical College, Changde, Hunan province, China (N29°02'29.74", E111°38'05.31", 34 m). The voucher specimens were well placed at the College of Life and Environmental Sciences, Hunan University of Arts and Sciences (Contact Person: Kerui Huang, huangkerui008@163.com, voucher number JZH007).

The library was constructed using the DNeasy plant tissue kit (TIANGEN Biotech Co., Ltd., Beijing) and the sequencing was performed on an Illumina HiSeq 2500 platform (San Diego, CA), both outsourced to Shanghai Personalbio Technology Co., Ltd. (China)

Chloroplast Genome Assembling and Annotation

After filtering out the low-quality reads using fastp [14], 81,419,412 clean reads were retained for further analysis. The chloroplast genome of C. officinalis was de novo assembled using GetOrganelle v1.7.5 [15] with parameters set as -R 15 -k 21,45,65,85,105 -F embplant_pt. Subsequently, the assembled chloroplast genome was annotated using CPGAVAS2 [16] with default settings, and a circular genome map was visualized using CPGView (http://www.1kmpg.cn/cpgview/).

Phylogenetic Analysis

A total of 44 chloroplast genomes closely related to C. officinalis, along with 2 outgroups, were downloaded from GenBank for phylogenetic analysis. Among them, 74 protein-coding genes shared by all genomes were screened out for subsequent analysis. Sequence alignment of each gene was performed separately using MAFFT v7.313 [17]. Gblocks 0.91b was then utilized to remove poorly aligned regions of each gene. The filtered gene sequences were concatenated head-to-tail into supergenes [18]. Maximum likelihood phylogenies were generated using IQ-TREE v1.6.12 [19]. The TVM + F + I + G4 model was selected based on the Bayesian Information Criterion (BIC) in ModelFinder. This process was further strengthened with 5000 ultrafast bootstrap replications for robust statistical support along with Shimodaira-Hasegawa-like approximate likelihood ratio test.

Codon usage bias analysis

In this study, the chloroplast genome sequences of *C. officinalis*, *C. arvensis*, and *O. ecklonis* were used to analyze codon usage bias. Coding sequences (CDS) were meticulously screened to meet specific criteria: multiples of three in base count, sequence length ≥ 300 bp, inclusion of only A, T, C, G bases, presence of start (ATG) and stop codons (TAG, TGA, TAA), and absence of internal stop codons and duplicate sequences, retaining 53 CDS for each species. Using CodonW and CUSP online software, metrics such as ENc, RSCU, CAI, CBI, Fop, and GC content were calculated. Codons for Met, Trp, and stop codons were excluded. Analyses including ENc-plot, PR2-plot, neutrality plot [21–22], and correspondence analysis based on RSCU values [23–24] were conducted, assessing the influence of mutation pressure and natural selection on codon usage bias.

Divergence time estimation

The divergence times for the species included in our phylogenetic analysis were estimated using the Markov chain Monte Carlo (MCMC) approach implemented in the PAML software package, specifically utilizing its MCMCtree program [25]. The optimal phylogenetic tree topology for our dataset was determined using IQ-TREE.

For calibrating the molecular clock, we incorporated three fossil-based calibration points derived from previous studies [25–32]. These calibration points were as follows: (F1) between 22.7 and 38.8 million years ago (Ma), (F2) between 17.4 and 44.7 Ma, and (F3) between 1.13 and 33.48 Ma. These points were strategically chosen to constrain each corresponding node in the phylogenetic tree.

Our analysis employed the independent rates model (IRM), which assumes a lognormal distribution for rate variation among lineages. The HKY85 model was selected for nucleotide substitution, with the alpha parameter for gamma-distributed rate variation across sites set at 0.5. The birth-death process model was used to establish priors for node ages within the phylogenetic tree. We adhered to the default settings for this model, with the parameters λ (birth rate) and μ (death rate) both set to 1, and the sampling proportion (s) set to 0. For the MCMC analysis, posterior probabilities of the parameters were estimated. The initial 10% of trees generated were discarded as burn-in to ensure sampling from a stationary distribution. Subsequent trees were sampled every 10 iterations, culminating in a total of 10,000 sampled trees for the final analysis.

Results

The chloroplast feature of C. officinalis

The chloroplast genome of *C. officinalis* (Fig. 1) is a circular molecule of 150,465 bp in length (Fig. 2), which consists of four parts: a large single-copy region (LSC) with a length of 83,056 bp; a small single-copy region (SSC) with a length of 17,911 bp; and two inverted repeat regions (IRs) of 24,749 bp, each. The G + C content of the whole chloroplast was 37.75%, and the IRs were 43.11%, which was higher than that of the LSC and the SSC regions (35.84% and 31.81%), The schematic representation of the entire chloroplast genome of *Calendula officinalis* is depicted in Figure S1, while Figure S2 illustrates the uniform mapping depth across the entire genome, indicating an absence of heteroplasmy. The genome contains 131 genes, including 86 protein-coding genes, eight rRNA genes, and 37 tRNA genes, the cis-splicing genes and trans-splicing gene rps12 in the chloroplast genome of *Calendula officinalis* can be found in Figure S3.

Phylogenetic analysis

Based on the chloroplast genome of *C. officinalis*, the Maximum-likelihood (ML) tree was constructed (Fig. 3) using 74 protein-coding genes, which helps to determine *C. officinalis*' phylogenetic status. The results (Fig. 3) show that *C. officinalis* and *C. arvensis* were within one clade with a support of 100%, and also, from a local point of view, the relationship between *C. officinalis*, *C. arvensis*, and *O. ecklonis* was very close, although *O. ecklonis* does not belong to the *Calendula* genus, which is consistent with the previous study [33]. Interestingly, *Crassocephalum crepidioides*, *Gynura japonica*, *Jacobaea vulgaris*, and *Seneico vulgaris* were found to be more closely related to *C. officinalis* in our study, compared to previous reports. This close relationship represents a novel finding, likely attributed to the differences in sequence data and phylogenetic methods employed in this study, as the protein-coding genes extracted from complete chloroplast genomes contain richer information compared to previous marker genes used.

Codon composition analysis

The condom composition for CDS of the three species (*Calendula officinalis*, *Calendula arvensis*, and *Osteospermum ecklonis*) was analyzed (Table 1), and the GC content of chloroplast-encoded genes in the three Asteraceae plants is 38.49%, 38.50%, and 38.54%, respectively. The GC content varies at different positions, with the first, second, and third positions in the codons all having a GC content below 50%. The highest content is at the first base, and the lowest content is at the third base, showing a trend of GC1 > GC2 > GC3. This indicates that the chloroplast genome sequences of the three Asteraceae plants are rich in A/T bases, particularly at the third position of the codons. the ENc values of chloroplasts in the three Asteraceae plants (*C. officinalis*, *C. arvensis*, and *O. ecklonis*) are 37.6959.17, 38.7459.17, and 39.1 ~ 58.98, with average values of 47.34, 47.41, and 47.65, respectively. All of these values are significantly greater than 35, indicating that the codon usage bias in the chloroplast genomes is relatively weak.

Table 1
GC content and effective number of codons (ENc) in *Calendula officinalis*, *Calendula arvensis*, and *Osteospermum ecklonis*.

Species	GeneBank ID	GCall	GC1	GC2	GC3	GC3s	ENc
Calendula officinalis	OP161555.1	38.49	47.18	39.96	28.34	25.21	47.34
Calendula arvensis	NC066014.1	38.50	47.28	39.97	28.27	25.15	47.41
Osteospermum ecklonis	NC061392.1	38.54	47.19	39.82	28.61	25.55	47.65

Note: GCall: Overall guanine-cytosine (GC) content in the species' genomic sequence (percentage); GC1: GC content at the first codon position in the species' genomic sequence (percentage); GC2: GC content at the second codon position in the species' genomic sequence (percentage); GC3: GC content at the third codon position in the species' genomic sequence (percentage); GC3s: GC content at synonymous third codon positions in the species' genomic sequence (percentage); ENc: Effective number of codons.

Codon usage bias analysis

By using GC3 and ENc as the X-axis and Y-axis, respectively, for the ENc-plot analysis, the influence of nucleotide composition on codon preference can be detected. When genes are distributed along the standard curve or near it, it indicates that the codon preference of the gene is affected only by mutations. However, when genes fall far below the standard curve, it indicates that the codon preference of the gene is affected by selection. From Fig. 4, it can be observed that the ENc-plot diagrams of the chloroplast genomes of the three Asteraceae plants (*C. officinalis*, *C. arvensis*, and *O. ecklonis*) are similar, some genes are distributed along the standard curve or close to it, indicating that their codon preference is mainly influenced by nucleotide mutations. However, some other genes deviate from the standard curve, suggesting that nucleotide mutations are not the main factor affecting their codon preference and that they may be affected by other factors such as natural selection. In addition to the similarity, there are a few differences, for example, photosynthesis-related genes of *C. officinalis* and *C. arvensis* are mainly concentrated near the standard curve, while those of *O. ecklonis* deviate below slightly, indicating that the photosynthesis-related genes of *O. ecklonis* might be more influenced by selection, while those of *C. officinalis* and *C. arvensis* are primarily affected by mutations

Parity rule 2 plots (PR2 plot) were generated respectively for the three Asteraceae plants (*C. officinalis*, *C. arvensis*, and *O. ecklonis*) using the chloroplast's protein-coding genes (Fig. 5). It can be easily noticed that all three plots were with great similarity. Firstly, the majority of their coordinate points are not uniformly distributed across the four regions but are mainly concentrated in the region where G3/(G3 + C3) > 0.5 and A3/(A3 + T3) < 0.5. Then, genes of the small subunit of ribosome of all three species tend to use A more, while Photosynthesis-related genes lean towards using T. However, overall, the usage frequency of the third base T in the codon is higher than A, and the usage frequency of G is higher than C. If codon usage bias were solely caused by nucleotide mutations, the usage frequencies of A/T and G/C should be equal. Therefore, the PR2-plot analysis results, combined with the ENc-plot analysis, indicate that the codon usage bias in the chloroplast genomes of the three Asteraceae plants is formed by the

combined effects of nucleotide mutations and natural selection. The similarity of the PR2-plot analysis results reflects the similarity of their phylogenetic relationships.

The correlation between codon GC12 and GC3 for the chloroplast genomes of the three Asteraceae plants (*C. officinalis*, *C. arvensis*, and *O. ecklonis*) is quite familiar as the Neutrality plot showed (Fig. 6). The codon GC12 values of the three plants' chloroplast genomes are distributed between 27.85 and 58.88, while GC3 values are distributed between 18.40 and 36.74, indicating that the frequency of using A/T at the third codon position is higher than G/C. The slope of the regression line fitted with GC12 and GC3 ranges from 0.13 to 0.22, with $R^2 > 0$, suggesting a positive correlation between G12 and G3 values. However, the two-tailed test did not reach significant levels ($P > 0.05$) for all three species, indicating that the mutation patterns of the first and second bases are different from the third base, and the codon usage bias is more affected by natural selection than by nucleotide mutations. Additionally, the regression coefficient of *O. ecklonis* is closest to 0, indicating that its chloroplast genome codon preference is most influenced by natural selection, while *C. officinalis* has the furthest regression coefficient from 0, suggesting that its chloroplast genome codon preference is least influenced by natural selection compared to the other two Asteraceae plants.

As the result of the correspondence analysis, the genes (CDS sequences) of the chloroplast genomes of the three Asteraceae plants (*C. officinalis*, *C. arvensis*, and *O. ecklonis*) are distributed on the figure with the first major factor axis as the x-coordinate and the second major factor axis as the y-coordinate (Fig. 7). The origin represents the average RSCU values of all genes relative to the first axis and the second axis. The sum of the proportions of the total variation accounted for by the first four principal factor axes in the three Asteraceae plants are 34.80%, 34.69%, and 33.66%, respectively. The proportion of the total variation accounted for by the first principal factor axis is 9.93%, 9.93%, and 9.23%, respectively, indicating that the first axis contributes the most to the variation, and the contribution of the remaining factor axes decreases successively. This again suggests that the formation of codon usage bias characteristics in the chloroplast genes of the three Asteraceae plants is not influenced by a single factor but is the result of the combined action of multiple factors.

To explore the factors affecting the distribution of each gene in the correspondence analysis plot of the chloroplast genomes of the three Asteraceae plants, correlation analysis between the first axis and GC3s, ENc, CAI, CBI, and Fop, respectively was performed. As can be seen from Table 2, the GC3s and CAI values of *C. officinalis* are significantly correlated with the first axis ($P < 0.05$); the GC3s and CAI values of *C. arvensis* are also highly significantly correlated with the first axis ($P < 0.05$), and the ENc value is significantly correlated with the first axis; the ENc value of *O. ecklonis* is significantly correlated with the first axis. It can be found that, when not considering the slope direction, the correlation between the first axis and various indicators of *C. officinalis* and *C. arvensis* is closer, while there is a larger difference with *O. ecklonis*. This pattern is consistent with that shown in the phylogenetic tree, reflecting the similarity and differences among the three, which indicate that correspondence analysis may reveal the commonalities and subtle differences in codon usage bias among the three species, which may be an

important characteristic reflecting the differences in their phylogenetic relationships, even if their relationships are relatively close.

Table 2
Correlations of ENc, GC3s, GC, CAI, and Fop with the first axis of correspondence analysis for *Calendula officinalis*, *Calendula arvensis*, and *Osteospermum ecklonis* respectively.

Features	*Calendula officinalis*	*Calendula arvensis*	*Osteospermum ecklonis*
ENc	-0.173694761	0.339292696*	-0.298674522*
GC3s	-0.355259062*	0.386953196**	0.148086675
GC	0.027698715	-0.051792803	0.117641069
CAI	0.323028467*	-0.366083451**	-0.016358954
Fop	0.174586161	-0.197259938	0.062726022

Note: The numerical values in the table represent correlation coefficients between the respective indices (ENc, GC3s, GC, CAI, and Fop) and the first axis of the corresponding analysis. An asterisk (*) indicates a significant correlation at $P < 0.05$, while two asterisks (**) denote a highly significant correlation at $P < 0.01$. ENc: Effective Number of Codons; GC3s: GC content at synonymous third codon positions in a species' genomic sequence (percentage); GC: Guanine-Cytosine content; CAI: Codon Adaptation Index; Fop: Frequency of Optimal Codons.

Our divergence time analysis, as illustrated in Fig. 8, indicates that the divergence of *C. officinalis* took place approximately 0.25 million years ago (Mya), situating it in the recent Quaternary period. Additionally, the genus *Calendula* is estimated to have originated around 2.38 Mya, also during the Quaternary. Furthermore, the common ancestor of *Calendula* and *Osteospermum* is traced back to roughly 18.77 Mya, placing this divergence squarely within the Miocene epoch of the Neogene period, an era known for significant environmental and climatic shifts that likely influenced their evolutionary paths.

Discussion

C. officinalis is a versatile plant with applications in various fields, including ornamentation, chemistry, and pharmacology. Despite its widespread use, its chloroplast genome structure, features, phylogeny, and patterns of evolution and mutation have remained largely unexplored. This study aims to address this knowledge gap by examining the chloroplast genome, phylogeny, and codon usage bias of *C. officinalis*, thereby enhancing our understanding of its evolution, adaptation, and potential uses.

The chloroplast genome of *C. officinalis* was found to be a 150,465 bp circular molecule, containing a large single-copy region (LSC), a small single-copy region (SSC), and two inverted repeat regions (IRs). The genome's G + C content is 37.75%, and it comprises 131 genes, including 86 protein coding, eight rRNA, and 37 tRNA genes. A Maximum-likelihood (ML) tree was constructed using 74 protein-coding genes to establish the phylogenetic status of *C. officinalis*. Phylogenetic analysis revealed that *C. officinalis* and *C. arvensis* form a clade with 100% support, and their relationship with *O. ecklonis* is close, in accordance with a previous study. The analysis also indicated a closer relationship between *C.*

officinalis and four other species than previously reported (Fu et al. 2016). This discrepancy could be attributed to differences in sequences and methods employed for phylogenetic analysis, warranting further investigation.

Codon usage bias is a vital element of evolution across diverse genomes, which is influenced by multiple biological factors including gene expression, gene length, tRNA abundance, mutation bias, and GC composition, as evidenced by a wealth of studies [34–40]. Nevertheless, it's the interplay between directional mutation pressure and natural selection that primarily governs codon usage bias across diverse organisms, forming the bedrock of interspecies and intragenomic codon usage disparities [41]. Plant genomes further demonstrate the complexity of codon usage bias; the nuclear gene codon preference is largely shaped by nuclear acid composition constraints, whereas in the realm of chloroplast and mitochondrial genomes, natural selection takes precedence [42–43]. The effective number of codons (ENc) is a common metric to quantify the degree of deviation in codon usage from random selection. Ranging in value from 20 to 61, ENc helps evaluate the strength of codon usage bias in genomes or genes. Smaller ENc values signify stronger codon preference, while larger values indicate weaker codon preference. Notably, when the ENc value is less than or equal to 35, the codon usage bias phenomenon is considered more significant.

In our research, we found that the chloroplast genome sequences of the three Asteraceae plants *C. officinalis*, *C. arvensis*, and *O. ecklonis* were rich in A/T bases, with ENc values ranging from 37.69 to 59.17, 38.74 to 59.17, and 39.1 to 58.98, and average values of 47.34, 47.41, and 47.65, respectively. As all these values are significantly greater than 35, this suggests that the codon usage bias in the chloroplast genomes of these species is relatively weak. Further analysis using ENc-plot (Fig. 4), PR2-plot (Fig. 5), Neutrality plot (Fig. 6), and correspondence analysis (Fig. 7) revealed that the codon usage bias in *C. officinalis*, *C. arvensis*, and *O. ecklonis* is a result of the combined effects of natural mutation and selection pressure. In addition, the codon usage bias patterns in these three species are highly similar, providing a robust explanation for their phylogenetic similarities. This finding suggests that these species may have been subjected to similar environmental conditions and selection pressures during their evolutionary process. This similarity in environmental conditions and selection pressures could also account for the close phylogenetic relationship between the species *O. ecklonis* and *C. officinalis* and *C. arvensis*. Moreover, we discovered that the correlation of the first axis of the correspondence analysis with GC3s, ENc, CAI, CBI, and Fop can effectively reflect the subtle differences in codon usage bias patterns among the three species. Interestingly, these differences are consistent with the results of the phylogenetic tree, indicating that this phenomenon is worthy of further study.

The divergence of the genus *Calendula* around 2.38 Ma (Fig. 8), within the Quaternary period, corresponds to a phase of Earth's history marked by intense climatic fluctuations. This period, characterized by repeated glacial and interglacial cycles [43–44], would have imposed strong selective pressures on plant species, driving adaptive responses. The speciation of *C. officinalis* during this time suggests its evolutionary resilience and adaptability to changing environments. This aligns with our

findings of weak codon usage bias in the chloroplast genome, indicative of a balanced selection-mutation dynamic possibly influenced by these environmental shifts.

The emergence of the common ancestor of *Calendula* and *Osteospermum* around 18.77 Ma (Fig. 8) in the Miocene epoch of the Neogene period coincides with significant global climate changes from warmer to cooler conditions [44]. The Miocene epoch, known for its extensive tectonic activities and consequent ecological shifts [45], likely provided diverse niches and selective pressures that catalyzed speciation events. The similarity in codon usage bias patterns among *Calendula* and *Osteospermum* species provides further evidence of their phylogenetic relationship and shared evolutionary history. This similarity suggests a parallel adaptation route, possibly as a response to similar environmental pressures over time.

The insights gleaned from this research not only improve our understanding of the evolutionary relationships and adaptation mechanisms of *C. officinalis* and related species but also lay the groundwork for future investigations into the potential applications of these plants. Understanding the divergence time and evolutionary adaptations of *C. officinalis* opens avenues for exploring its potential applications in pharmacology and agriculture. Future research could delve deeper into how specific adaptations in the chloroplast genome have contributed to its medicinal and ornamental properties.

Conclusion

In summary, our research provides a comprehensive analysis of the chloroplast genome, phylogenetic relationships, codon usage bias, and divergence time of *C. officinalis*. It highlights the close evolutionary kinship between *C. officinalis*, *C. arvensis*, and *O. ecklonis*, supported by similarities in codon usage patterns and divergence timelines. These findings suggest that shared environmental selection pressures have played a significant role in their evolutionary paths. The study not only advances our understanding of *C. officinalis*' evolution and adaptation but also lays the groundwork for future exploration in the Asteraceae family.

Declarations

Conflicts of Interest

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.

Author Contributions

Y.W. and K.H. conceived and designed the study. N.Z., Y.W., K.H., P.X., A.D., M.J., and P.M. collected the samples. K.H. and N.Z. performed genome assembling and data analysis. N.Z. drafted the manuscript. N.Z., H.Y., R.H., J.L., F.H., and H.H. revised the manuscript. All authors discussed the results, critically reviewed the manuscript, and approved the final version.

Statement of Permission and Compliance

We confirm that *Calendula officinalis* materials used in this study were collected with permission from the parterre of Changde Vocational Technical College. The collection complied with all relevant local legislation, and appropriate permissions were granted before the samples were collected.

All experiments and field studies on *Calendula officinalis* in this research complied with local legislation and were carried out in accordance with relevant institutional, national, and international guidelines and legislation. The experimental protocols were approved by the relevant ethics committee.

Funding

This work was supported by the Natural Science Foundation of Hunan Province (2023JJ30436, 2022JJ50249 and 2022JJ40291), the Scientific Research Foundation of Hunan Provincial Education Department (22A0487), Key Research Project of Hunan University of Arts and Science (E06022005), and the Scientific Research Youth Foundation of Education Department of Hunan Province (21B0610).

Acknowledgments

We thank Jun Yan and Mi He for their assistance with the phylogenetic analysis.

Data Availability Statement

The complete chloroplast genome sequence of *C. officinalis* has been deposited in the GenBank database under the accession number OP161555 (https://doi.org/www.ncbi.nlm.nih.gov/nuccore/OP161555). The associated BioProject and Bio-Sample numbers are PRJNA1019102 and SAMN37474090, respectively.

References

1. Bayat, H., Alirezaie, M. & Neamati, H. Impact of exogenous salicylic acid on growth and ornamental characteristics of calendula (*Calendula officinalis* L.) under salinity stress. *J. Stress Physiol. Biochem.* **8,** 258-267 (2012).
2. Jan, N., Andrabi, K.I. & John, R. Calendula officinalis-an important medicinal plant with potential biological properties. *Proc Indian Natn Sci Acad* **83,** 769-787 (2017).
3. Ashwlayan, V.D., Kumar, A. & Verma, M. Therapeutic potential of Calendula officinalis. *Pharm Pharmacol Int J* **6,** 149-155 (2018).
4. Green, B.R. Chloroplast genomes of photosynthetic eukaryotes. *Plant J.* **66,** 34-44 (2011).
5. Sugiura, M., Shinozaki, K., Zaita, N., Kusuda, M. & Kumano, M. Clone bank of the tobacco (*Nicotiana tabacum*) chloroplast genome as a set of overlapping restriction endonuclease fragments: mapping of eleven ribosomal protein genes. *Plant Sci.* **44,** 211-217 (1986).
6. Sugiura, M. The chloroplast genome. *Plant Mol. Biol.* **19,** 149–168 (1992).

7. Buhr, F. et al. Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations. *Mol. Cell* **61,** 341-351 (2016).
8. Zhou, Z. et al. Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proc. Natl. Acad. Sci.* **113,** E6117-E6125 (2016).
9. Peden, J.F. Analysis of codon usage. *BioSyst.* **5,** 73-74 (2000).
10. Sharp, P.M., Stenico, M., Peden, J.F. & Lloyd, A.T. Codon usage: mutational bias, translational selection, or both? *Biochem. Soc. Trans.* **21,** 835-841 (1993).
11. Subramanian, S. Nearly neutrality and the evolution of codon usage bias in eukaryotic genomes. *Genetics* **178,** 2429-2432 (2008).
12. Qin, H., Wu, W.B., Comeron, J.M., Kreitman, M. & Li, W.H. Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. *Genetics* **168,** 2245-2260 (2004).
13. Xing, Z.B., Cao, L., Zhou, M. & Xiu, L.S. Analysis on codon usage of chloroplast genome of Eleutherococcus senticosus. *Chin J Chin Mater Med* **38,** 661-665 (2013).
14. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34,** i884-i890 (2018).
15. Jin, J.J., Yu, W.B., Song, Y., dePamphilis, C.W. & Yi, T.S. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* **21,** 241 (2020).
16. Shi, L.C. et al. CPGAVAS2, an integrated plastome sequence annotator and analyzer. **Nucleic Acids Res 47,** W65-W73 (2019).
17. Rozewicki, J., Li, S., Amada, K.M., Standley, D.M. & Katoh, K. MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Res.* **47,** W5-W10 (2019).
18. Guo, S. et al. A Comparative Analysis of the Chloroplast Genomes of Four Polygonum Medicinal Plants. *Front. Genet.* **13,** 764534 (2022).
19. Nguyen, L.T., Schmidt, H.A., von Haeseler, A. & Quang Minh, B. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32,** 268-274 (2015).
20. Sueoka, N. Translation-coupled violation of Parity Rule 2 in human genes is not the cause of heterogeneity of the DNA G+C content of third codon position. *Gene* **238,** 53-58 (1999).
21. Wei, L. et al. Analysis of codon usage bias of mitochondrial genome in Bombyx mori and its relation to evolution. *BMC Evol. Biol.* **14,** 1-12 (2014).
22. Wen, Y., Zou, Z., Li, H., Xiang, Z. & He, N. Analysis of codon usage patterns in Morus notabilis based on genome and transcriptome data. *Genome* **60,** 473-484 (2017).
23. James, F.C. & McCulloch, C.E. Multivariate analysis in ecology and systematics: panacea or Pandora's box? *Annu. Rev. Ecol. Syst.* **21,** 129-166 (1990).
24. Wang, Z. et al. Comparative analysis of codon usage patterns in chloroplast genomes of six Euphorbiaceae species. *PeerJ* **8,** e8251 (2020).

25. Puttick, M. N.. MCMCtreeR: functions to prepare MCMCtree analyses and visualize posterior ages on trees. *Bioinformatics*, **35**(24): 5321-5322.(2019).
26. Li, H. T. et al. Origin of angiosperms and the puzzle of the Jurassic gap. *Nature plants*, **5**(5), 461-470 (2019).
27. Kim, K. J., Choi, K. S., Jansen, R. K. Two chloroplast DNA inversions originated simultaneously during the early evolution of the sunflower family (Asteraceae). *Molecular biology and evolution*, **22**(9), 1783-1792 (2005).
28. Mandel, J. R. et al. A fully resolved backbone phylogeny reveals numerous dispersals and explosive diversifications throughout the history of Asteraceae. *Proceedings of the National Academy of Sciences*, **116**(28), 14083-14088 (2019).
29. Zhang, C. et al. Phylotranscriptomic insights into Asteraceae diversity, polyploidy, and morphological innovation. *Journal of Integrative Plant Biology*, **63**(7), 1273-1293 (2021).
30. Zhang, Q. et al. New insights into the formation of biodiversity hotspots of the Kenyan flora. *Diversity and Distributions*, **28**(12), 2696-2711 (2022).
31. Verboom, G. A., Stock, W. D., & Cramer, M. D. Specialization to extremely low-nutrient soils limits the nutritional adaptability of plant lineages. *The American Naturalist*, **189**(6), 684-699 (2017).
32. Foster, C.S.P. et al. Evaluating the impact of genomic data and priors on Bayesian estimates of the angiosperm evolutionary timescale. *Systematic Biology*, **66**(3), 338-351 (2017).
33. Fu, Z.X., Jiao, B.H., Nie, B., Zhang, G.J. & Gao, T.G. A comprehensive generic-level phylogeny of the sunflower family: Implications for the systematics of Chinese Asteraceae. *J. Syst. Evol.* **54,** 416-437 (2016).
34. Wang, B., Yuan, J., Liu, J., Jin, L. & Chen, J.Q. Codon usage bias and determining forces in green plant mitochondrial genomes. *J. Integr. Plant Biol.* **53,** 324-334 (2011).
35. Blake, W.J., Kaern, M., Cantor, C.R. & Collins, J.J. Noise in eukaryotic gene expression. *Nature* **422,** 633-637 (2003).
36. Ingvarsson, P.K. Gene expression and protein length influence codon usage and rates of sequence evolution in Populus tremula. *Mol. Biol. Evol.* **24,** 836-844 (2007).
37. Duret, L. & Mouchiroud, D. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc. Natl. Acad. Sci.* **96,** 4482-4487 (1999).
38. Rao, Y., Wu, G., Wang, Z., Chai, X., Nie, Q. & Zhang, X. Mutation bias is the driving force of codon usage in the *Gallus gallus* genome. *DNA Res.* **18,** 499-512 (2011).
39. Sueoka, N. & Kawanishi, Y. DNA G+C content of the third codon position and codon usage biases of human genes. *Gene* **261,** 53-62 (2000).
40. Wan, X.F., Xu, D., Kleinhofs, A. & Zhou, J. Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. *BMC Evol. Biol.* **4,** 1-11 (2004).
41. Sharp, P.M., Emery, L.R. & Zeng, K. Forces that influence the evolution of codon bias. *Philos T. R. Soc B* **365,** 1203-1212 (2010).

42. Liu, Q. & Xue, Q. Comparative studies on codon usage pattern of chloroplasts and their host nuclear genes in four plant species. *J. Genet.* **84,** 55-62 (2005).
43. Morton, B.R. & Wright, S.I. Selective constraints on codon usage of nuclear genes from *Arabidopsis thaliana*. *Mol. Biol. Evol.* **24,** 122-129 (2007).
44. Richter, C. et al. New insights into Southern Caucasian glacial–interglacial climate conditions inferred from Quaternary gastropod fauna. *Journal of Quaternary Science*, 35(5), 634-649 (2020).
45. Brown, S.C. et al. Persistent Quaternary climate refugia are hospices for biodiversity in the Anthropocene. *Nature Climate Change*, **10**(3), 244-248 (2020).
46. Holbourn, A.E. et al. Late Miocene climate cooling and intensification of southeast Asian winter monsoon. *Nature communications*, **9**(1), 1584 (2018).
47. Lin, C. et al. Himalayan Miocene adakitic rocks, a case study of the Mayum pluton: Insights into geodynamic processes within the subducted Indian continental lithosphere and Himalayan mid-Miocene tectonic regime transition. *Bulletin*, **133**(3-4), 591-611 (2021).

Figures



Figure 1

The picture of the collected sample of *Calendula officinalis*

The picture is self-taken nearby the Changde Vocational Technical College, Changde, Hunan province, China (N29°02'29.74", E111°38'05.31", 34 m). In the picture, *C. officinalis* is the plant 2-30 inches tall, with flowers usually 2-3 inches in diameter and a head-shaped inflorescence; the flowers are yellow to orange in color and hermaphrodite with tubular-shaped florets; The leaves are oblanceolate, alternate, sessile,

bright green, and usually about 4 inches long; Stems with umbel-like branches; as well as curved, ring-shaped, and sickle-shaped achene.

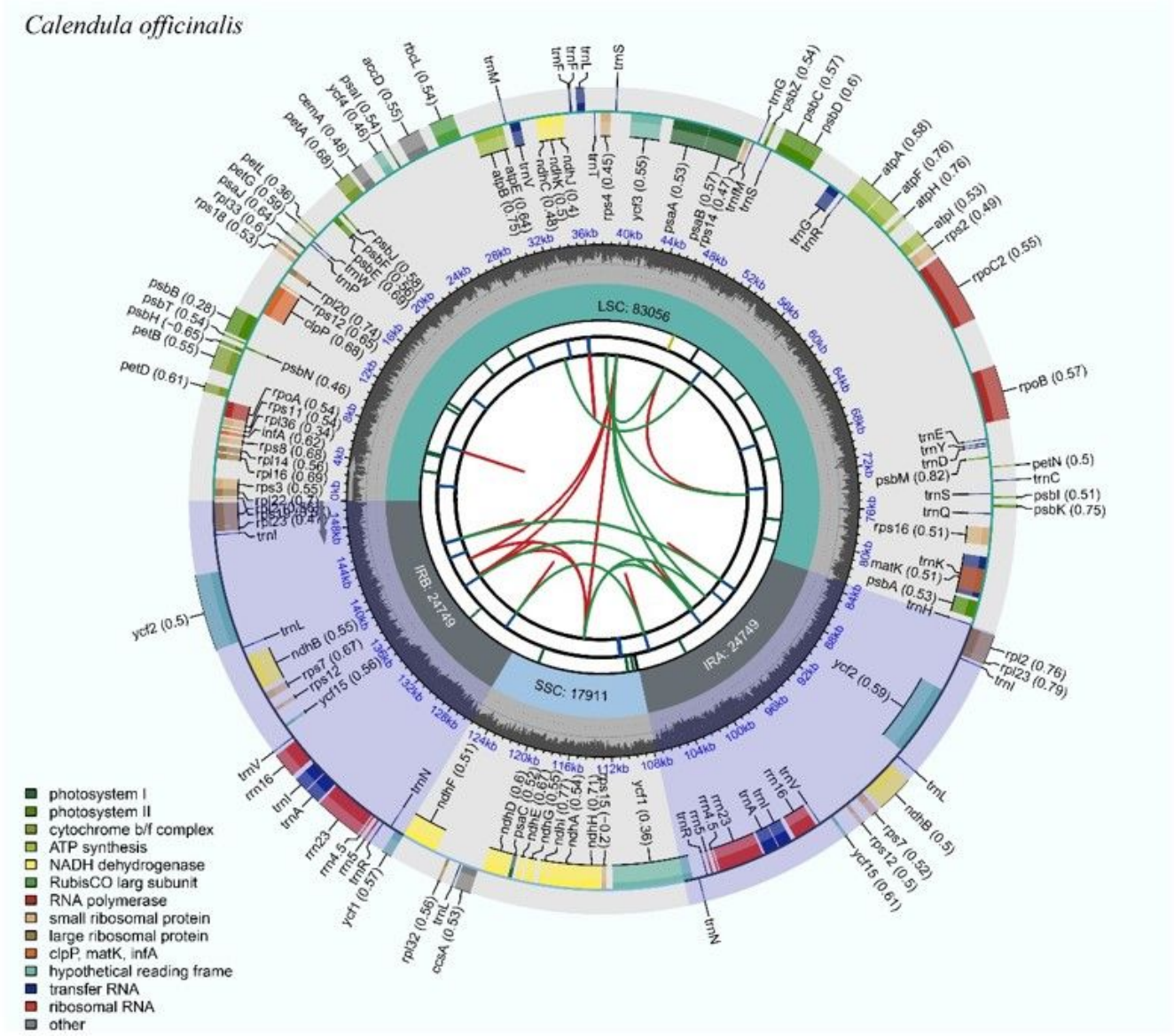


Figure 2

Gene map of the *Calendula officinalis*chloroplast genome.

The center of the figure represents the genome, and the various tracks surrounding it represent different features. The first track shows the dispersed repeats, while the second track displays the long tandem repeats as short blue bars. The third track indicates the short tandem repeats or microsatellite sequences as short bars with different colors. The fourth track shows the small single-copy (SSC), inverted repeat (Ira and Irb), and large single-copy (LSC) regions. The GC content along the genome is plotted on the fifth track. Finally, the genes are displayed on the sixth track.

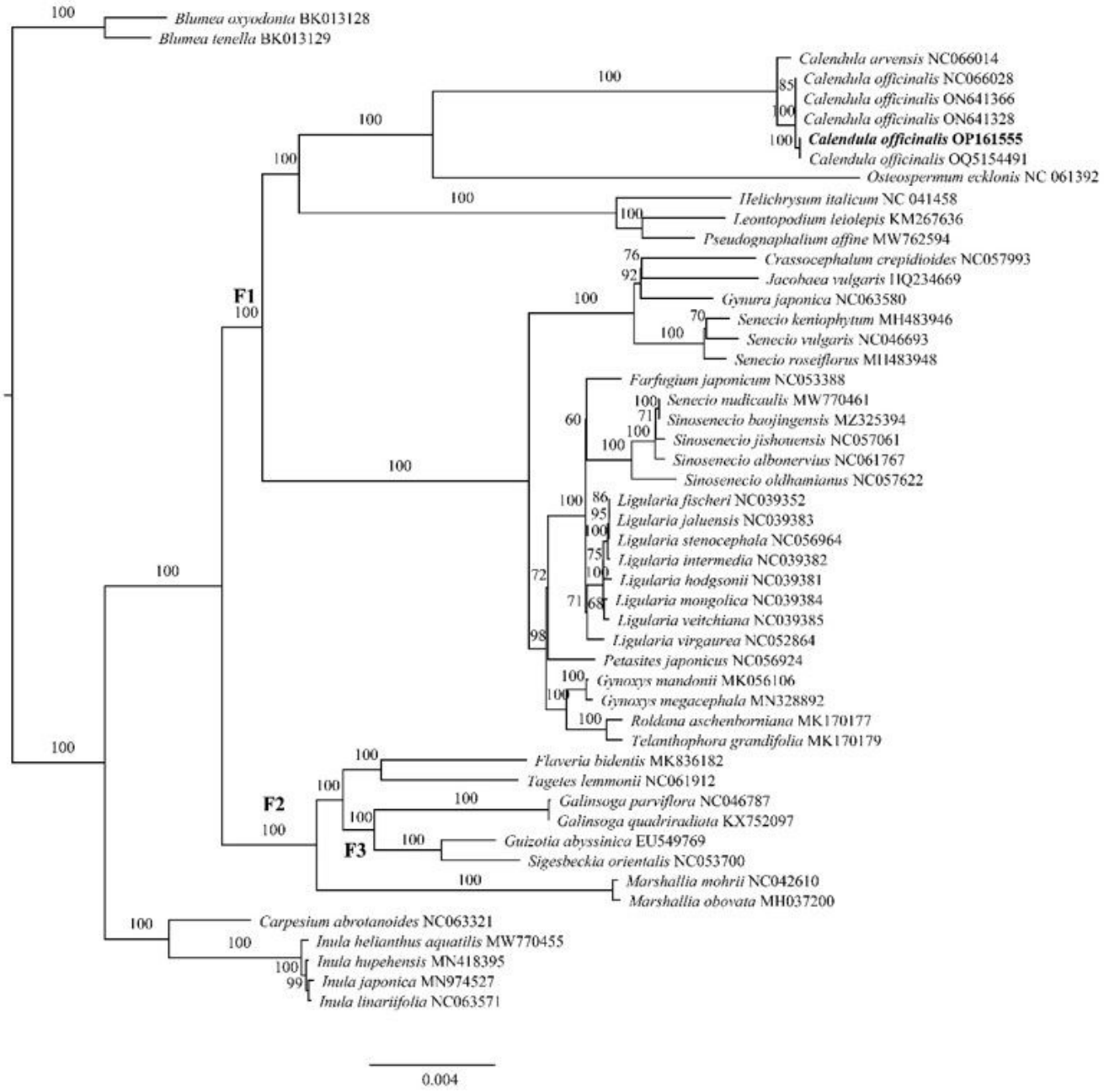


Figure 3

Maximum-likelihood (ML) tree of *Calendula officinalis* and 44 related species along with two outgroups using the IQ-Tree based on 74 protein-coding genes shared by all genomes. Bootstrap values are shown next to the nodes.

The following sequences were used: *Roldana aschenborniana* MK170177, *Telanthophora grandifolia* MK170179, *Crassocephalum crepidioides* NC057993, *Gynura japonica* NC063580 *Jacobaea vulgaris* HQ234669, *Senecio keniophytum*MH483946, *Senecio vulgaris* NC046693, *Senecio roseiflorus* MH483948, *Blumea oxyodonta* BK013128, *Blumea tenella* BK013129, *Carpesium abrotanoides* NC063321, *Inula helianthus aquatilis* MW770455, *Inula hupehensis* MN418395, *Inula japonica* MN974527, *Inula linariifolia* NC063571,*Flaveria bidentis* MK836182, *Tagetes lemmonii* NC061912, *Galinsoga parviflora* NC046787,

Galinsoga quadriradiata KX752097, *Guizotia abyssinica* EU549769, *Sigesbeckia orientalis* NC053700, *Marshallia mohrii* NC042610, *Marshallia obovata* MH037200, *Helichrysum italicum* NC041458, *Leontopodium leiolepis* KM267636, *Pseudognaphalium affine* MW762594, *Osteospermum ecklonis* NC061392, *Farfugium japonicum* NC053388, *Senecio nudicaulis* MW770461, *Sinosenecio baojingensis* MZ325394, *Sinosenecio jishouensis* NC057061, *Sinosenecio albonervius* NC061767, *Sinosenecio oldhamianus* NC057622, *Ligularia virgaurea* NC052864, *Ligularia fischeri* NC039352, *Ligularia stenocephala* NC056964, *Ligularia jaluensis* NC039383, *Ligularia intermedia* NC039382, *Ligularia mongolica* NC039384, *Ligularia veitchiana* NC039385, *Ligularia hodgsonii* NC039381, *Syneilesis aconitifolia* NC061374, *Petasites japonicus* NC056924, *Gynoxys mandonii* MK056106, *Gynoxys megacephala* MN328892. Calibration points F1, F2, and F3, used for divergence time analysis, are marked on the tree

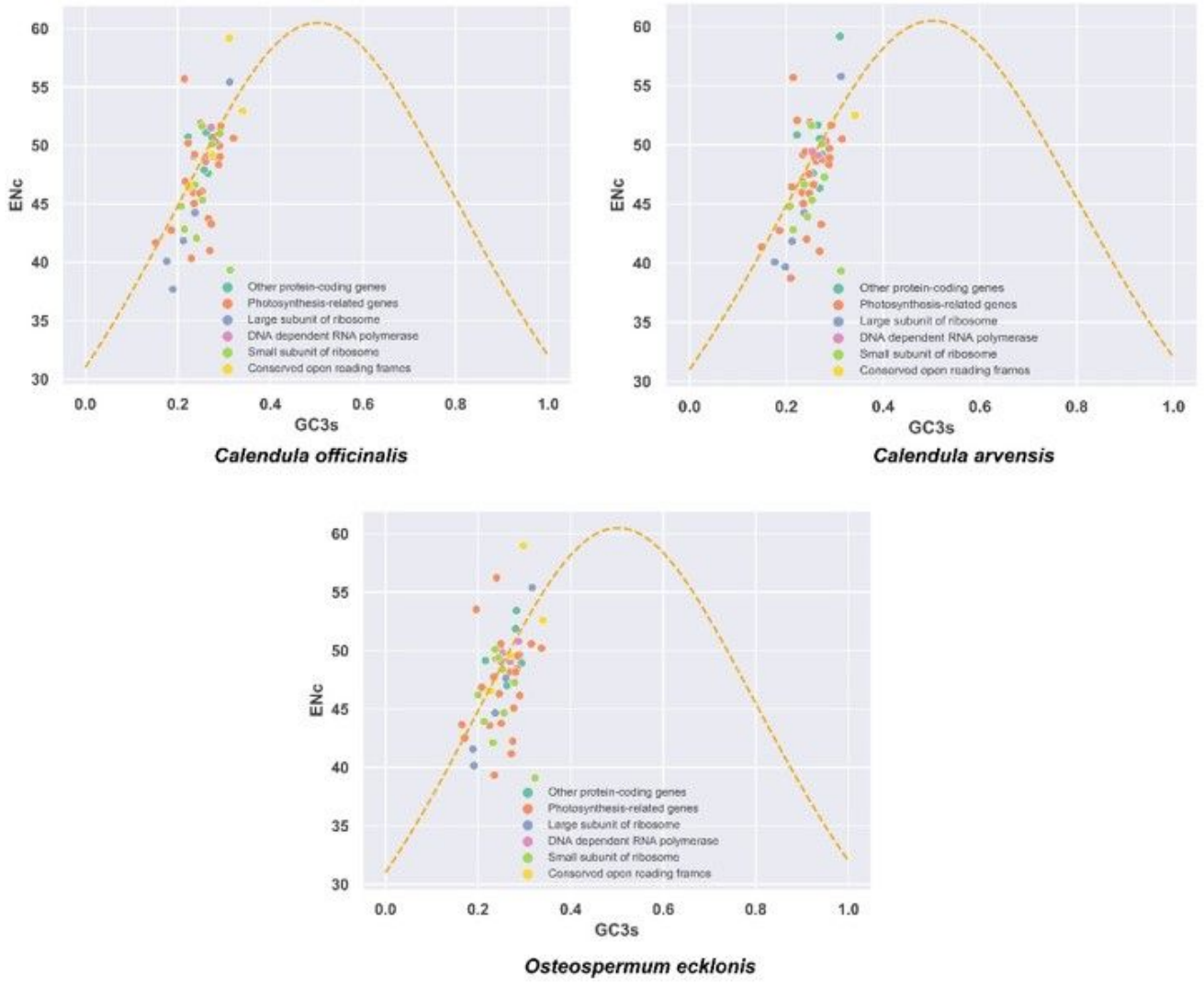


Figure 4

Effective number of codons (ENc) of each coding gene against GC3s for the chloroplast genome of *Calendula officinalis*, *Calendula arvensis*, and *Osteospermum ecklonis.*

The scatterplot displays the relationship between the effective number of codons (ENc) and GC content at the third codon position (GC3s) for each coding gene in the chloroplast genomes of the three species. Different colors represent genes of different types.

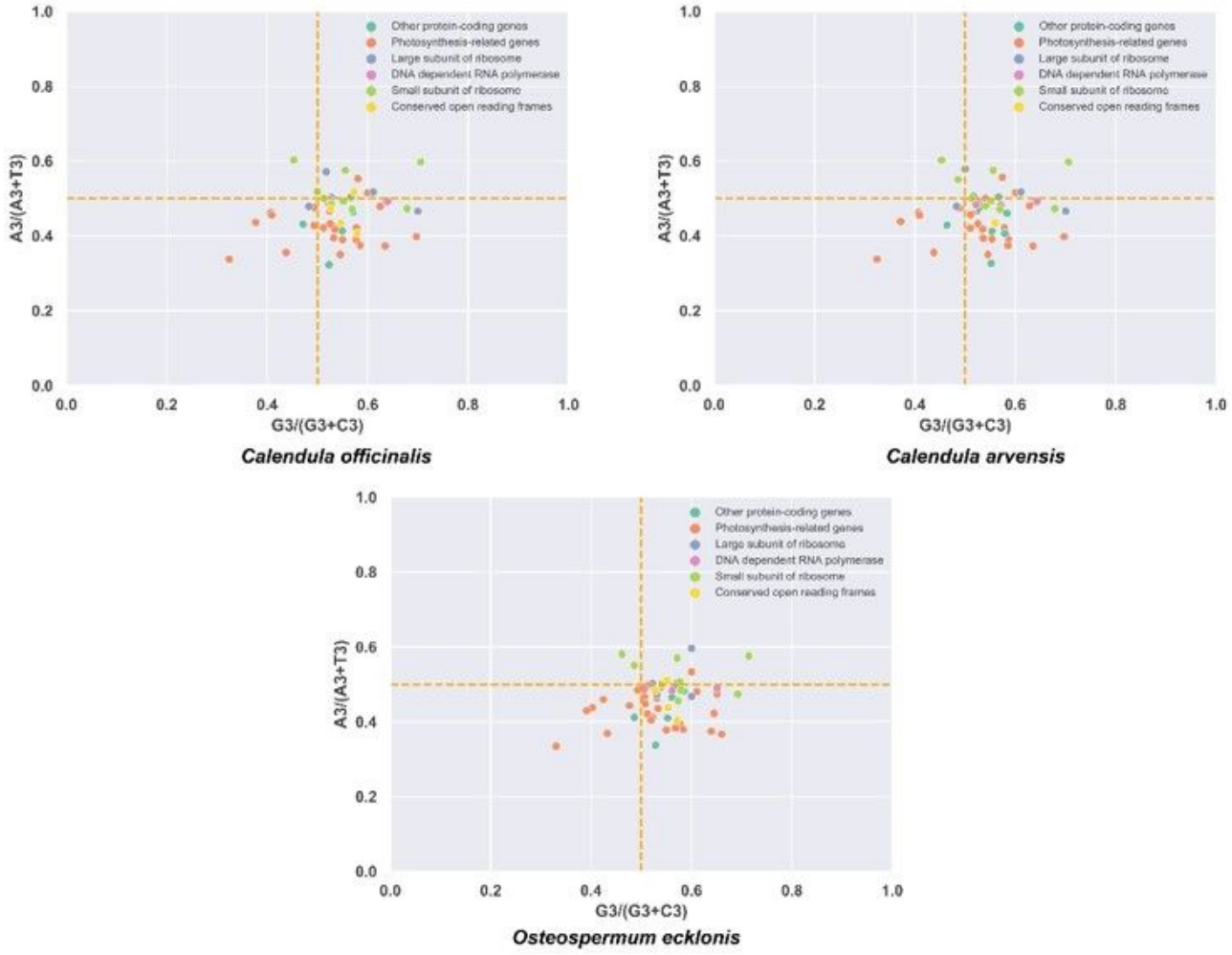


Figure 5

The parity rule 2 (PR2) bias plots for genes of the chloroplast genome of *Calendula officinalis*, *Calendula arvensis*, and *Osteospermum ecklonis.*

The scatterplot displays the relationship between A3/(A3+T3) and G3/(G3+C3) for each coding gene in the chloroplast genomes of the three species, representing the parity rule 2 (PR2) bias. Different colors represent genes of different types.

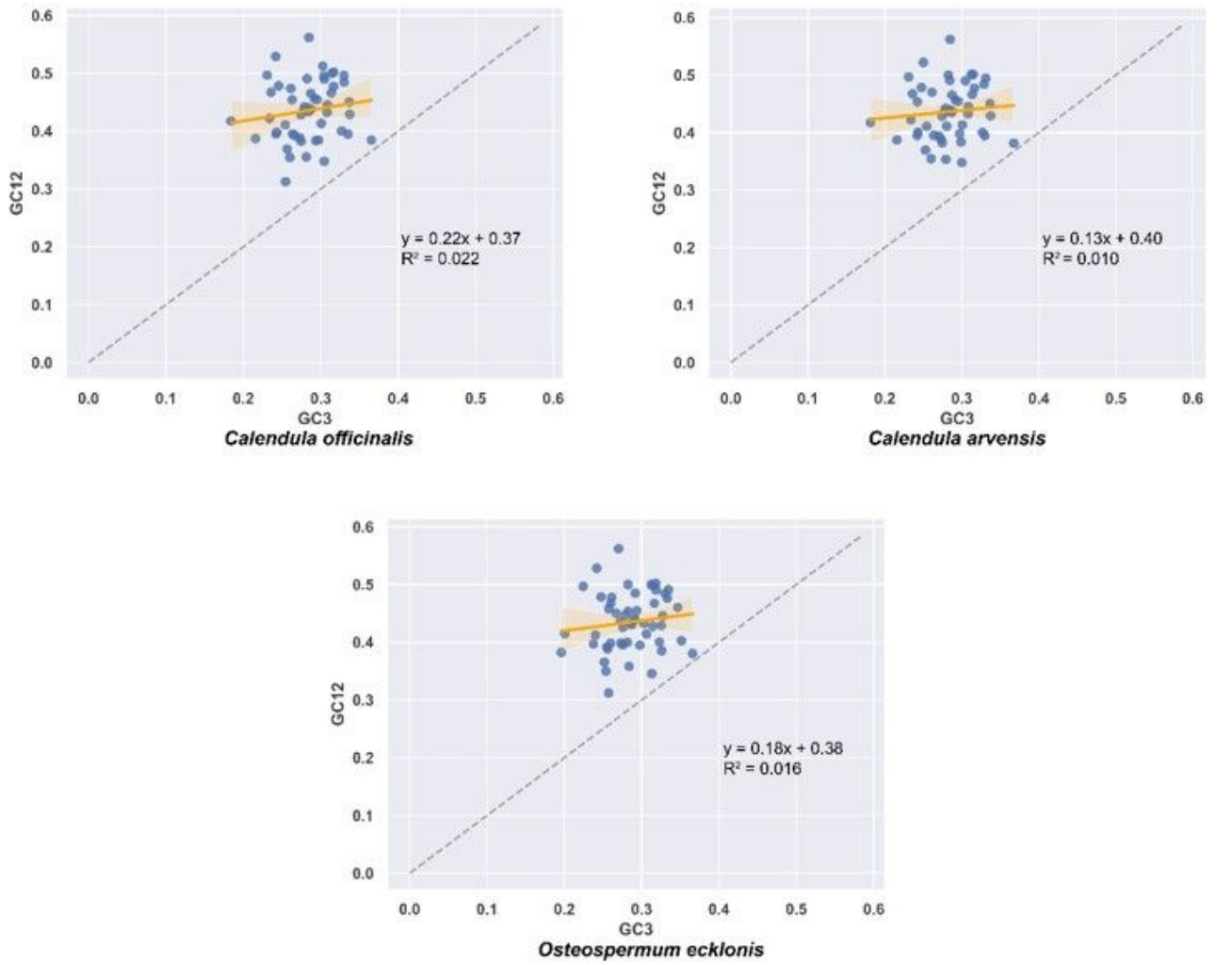


Figure 6

Neutrality plot for genes of the chloroplast genome of *Calendula officinalis*, *Calendula arvensis*, and *Osteospermum ecklonis.*

The scatterplot displays the correlation between the average GC content at the first and second codon positions (GC12) and the GC content at the third codon position (GC3) for each coding gene in the chloroplast genomes of the three species.

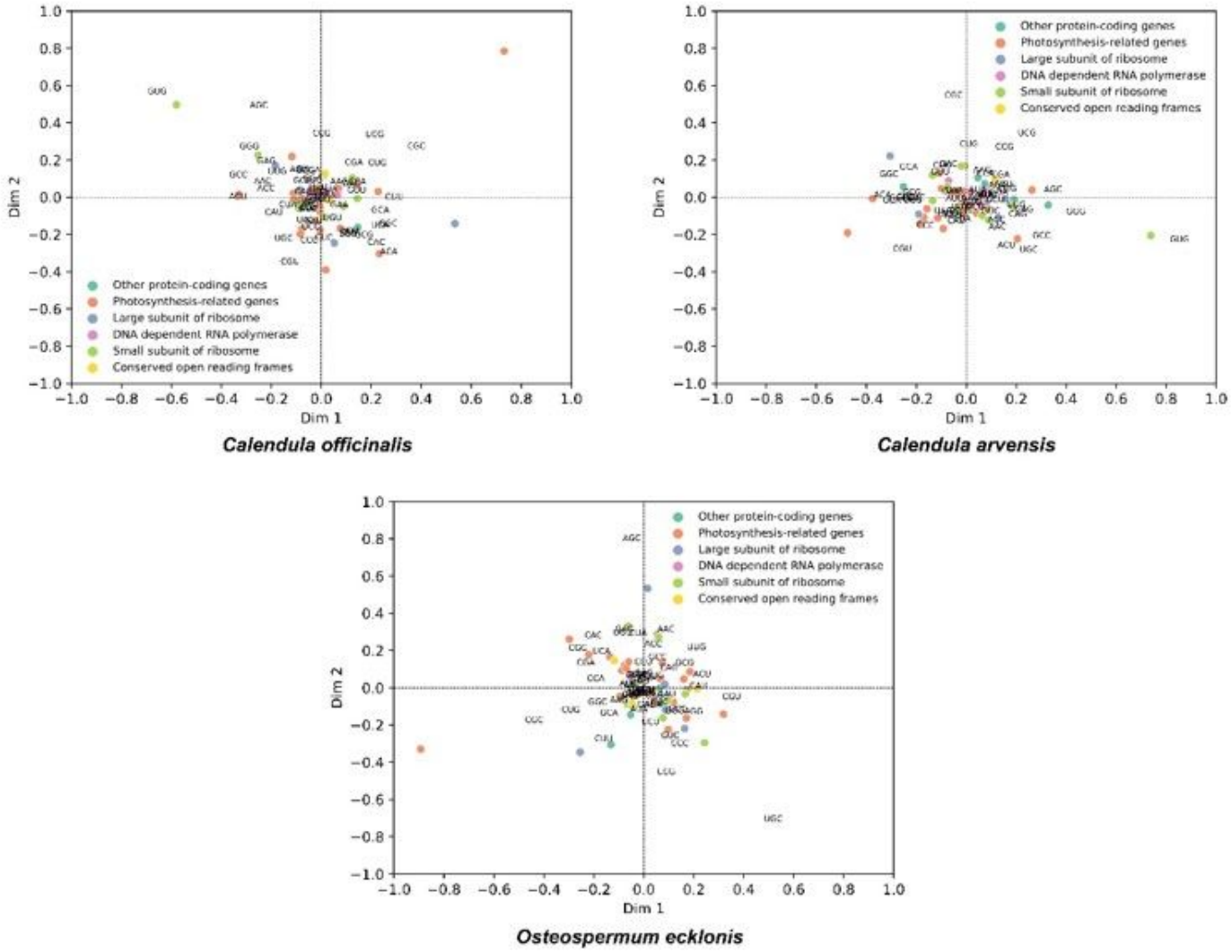


Figure 7

Correspondence analysis of the RSCU values of the chloroplast genome of *Calendula officinalis*, *Calendula arvensis*, and *Osteospermum ecklonis*.

Each point on the plot corresponds to the coordinates on the first and second-principal axes, representing the distribution of the genes based on their relative synonymous codon usage (RSCU) values. The codons also distribute in the plot.

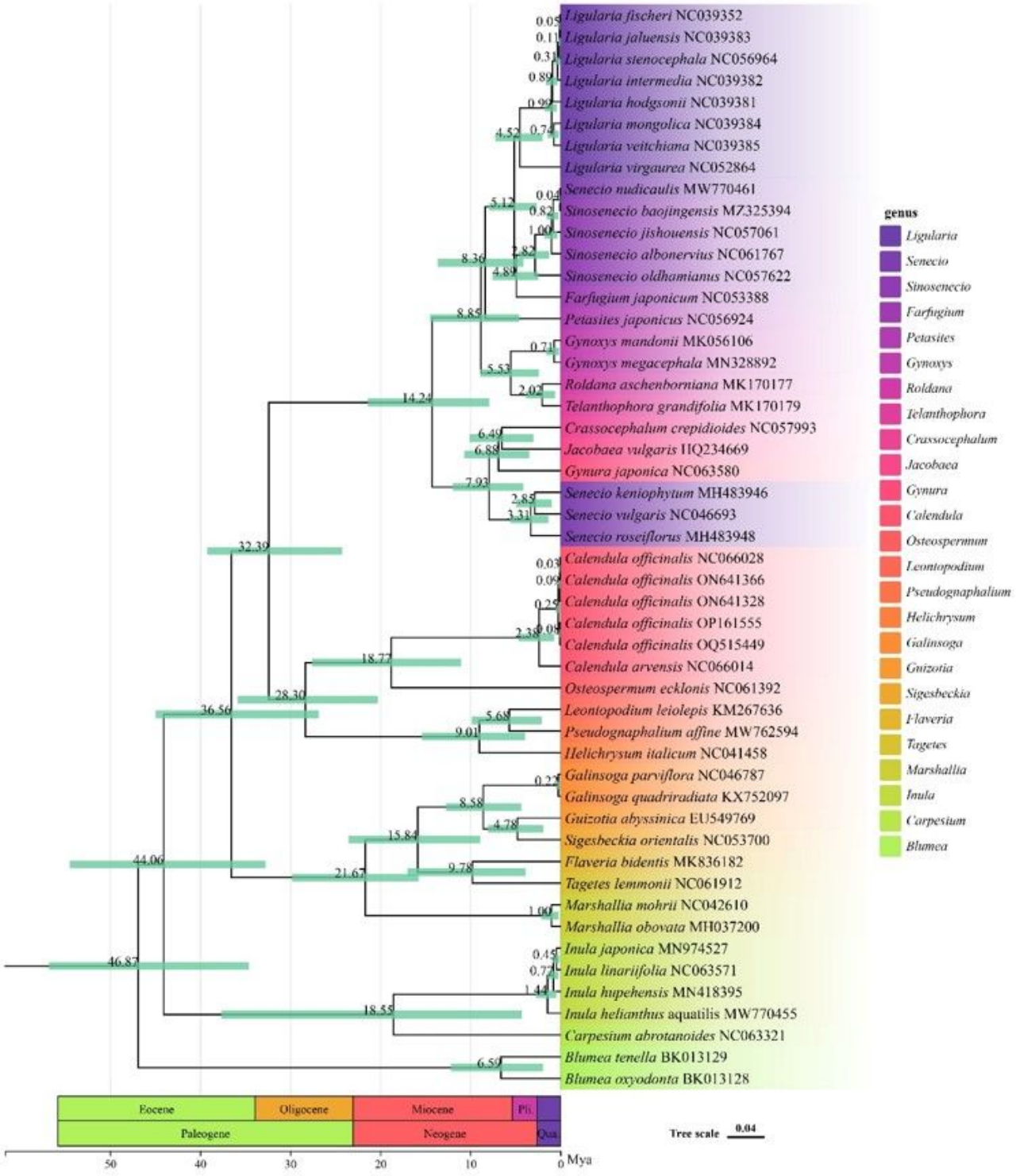


Figure 8

Divergence time estimation based on cp genome sequences.

The divergence times are exhibited on each node, whereas the greed bars represent the 95% highest posterior density interval for each node age.

Supplementary Files

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

- SupplementalMaterial.docx