frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 04 November 2024
DOI 10.3389/fpls.2024.1469375

OPEN ACCESS

EDITED BY
Baohua Wang,
Nantong University, China

REVIEWED BY
Xuming Li,
Hugo Biotechnologies Co., Ltd., China
Cao Deng,
DNA Stories Bioinformatics Center, China

*CORRESPONDENCE
Da-Wei Li
dli@dongyang-lab.org
Yang Dong
loyalyang@163.com
Sheng-Chang Duan
duanshengchang@163.com

†These authors have contributed equally to this work and share first authorship

RECEIVED 23 July 2024
ACCEPTED 14 October 2024
PUBLISHED 04 November 2024

CITATION
Chen B-Z, Yang Z-J, Yang L, Zhu Y-F, Li X-Z, Wang L, Zhou Y-P, Zhang G-H, Li D-W, Dong Y and Duan S-C (2024) Chromosome-scale genome assembly of *Codonopsis pilosula* and comparative genomic analyses shed light on its genome evolution. *Front. Plant Sci.* 15:1469375. doi: 10.3389/fpls.2024.1469375

Chromosome-scale genome assembly of *Codonopsis pilosula* and comparative genomic analyses shed light on its genome evolution

Bao-Zheng Chen[1,2†], Zi-Jiang Yang[3†], Ling Yang[4†], Yi-Fan Zhu[1,2], Xu-Zhen Li[2,5], Lei Wang[2], Ye-Peng Zhou[2], Guang-Hui Zhang[6], Da-Wei Li[2,5*], Yang Dong[2*] and Sheng-Chang Duan[2,5*]

[1]College of Food Science and Technology, Yunnan Agricultural University, Kunming, Yunnan, China, [2]Yunnan Provincial Key Laboratory of Biological Big Data, Yunnan Agricultural University, Kunming, Yunnan, China, [3]Bioinformatics Group, Wageningen University and Research, Wageningen, Netherlands, [4]Institute of Agro-Products of Processing and Design, Hainan Academy of Agricultural Sciences, Haikou, Hainan, China, [5]College of Plant Protection, Yunnan Agricultural University, Kunming, Yunnan, China, [6]National and Local Joint Engineering Research Center on Germplasm Innovation and Utilization of Chinese Medicinal Materials in Southwest China, Yunnan Agricultural University, Kunming, Yunnan, China

Introduction: *Codonopsis pilosula* is a significant plant in traditional Chinese medicine, valued for its edible and medicinal properties. However, the lack of available genomic resources has hindered further research.

Methods: This study presents the first chromosome-scale genome assembly of *C. pilosula* using PacBio CLR reads and Hi-C scaffolding technology. Additionally, Ks analysis and syntenic depth analysis were performed to elucidate its evolutionary history.

Results: The final assembly yielded a high-quality genome of 679.20 Mb, which was anchored to 8 pseudo-chromosomes with an anchoring rate of 96.5% and a scaffold N50 of 80.50 Mb. The genome assembly showed a high completeness of 97.6% based on Benchmarking with Universal Single-Copy Orthologs (BUSCO) analysis. Repetitive elements constituted approximately 76.8% of the genome, with long terminal repeat retrotransposons (LTRs) accounting for about 39.17%. Ks and syntenic depth analyses revealed that the polyploidization history of three platycodonoid clade species involved only the γ-WGT event. Karyotype evolutionary analysis identified an ancestral karyotype with 9 protochromosomes for the three platycodonoid clade species. Moreover, non-WGD genes, particularly those arising from tandem duplications, were found to contribute significantly to gene family expansion.

Discussion: These findings provide essential insights into the genetic diversity and evolutionary biology of *C. pilosula*, aiding its conservation and sustainable use.

KEYWORDS

***Codonopsis pilosula*, assembly, comparative genomics, gene family, genome evolution**

Introduction

The dried roots of *Codonopsis pilosula* (Franch.) Nannf. (Campanulaceae), referred to as "dang shen" in Chinese, are widely used in traditional Chinese medicine. *Codonopsis pilosula* is commonly used as a substitute for the more expensive *Panax ginseng* as a tonic agent, offering comparable therapeutic effects (College, 1986; Meng et al., 2020). Besides its significant medicinal value, *C. pilosula* is also used as a valuable vegetable plant, playing an important role in enhancing dietary nutrition and boosting overall health (He et al., 2016). Previous phytochemical analyses of *C. pilosula* have identified some important secondary metabolites, including triterpenoid saponins, phenylpropanoids, alkaloids, polyacetylenes, and other compounds (Lin et al., 2013; Jiang et al., 2016; Bai et al., 2020; Vo et al., 2024). These findings provide the basis for its pharmaceutical applications, which include anti-tumor (Xu et al., 2012), anti-inflammatory (Meng et al., 2020), immunomodulatory effects (He et al., 2016), anti-fatigue (Cai et al., 2014) and other biological activities (Zou et al., 2020).

Although many studies have contributed to the findings of *C. pilosula*, including the isolation of bioactive compounds (Bai et al., 2020), pharmacological research (Zou et al., 2020), molecular phylogenetic studies (Crowl et al., 2016), and more, the chromosome-level assembly of *C. pilosula* genome has not been reported. The majority of genetic information remains within the nuclear DNA, indicating that the genome-level evolutionary history of *C. pilosula* remains understudied. Furthermore, the genomes of two species within the Campanulaceae family, *C. lanceolata* (Jang et al., 2023) and *Platycodon grandiflorus* (Jia et al., 2022), have been deciphered, offering an excellent genetic resource to investigate their evolutionary relationships and genomic features. Additionally, the progress in genomics, transcriptomics, and related omics fields have greatly enhanced studies in evolutionary and conservation biology. For instance, leveraging one thousand plant transcriptomes, Leebens-Mack et al. (2019) established a strong phylogenomic framework to investigate the evolution of green plants, and their findings suggested that whole-genome duplications have repeatedly taken place during the evolution of flowering plants and ferns. Ma et al. (2022) conducted a study on *Acer yangbiense*, an endangered species with fragmented habitats and a restricted range in Yunnan, China. By resequencing the whole genomes of 105 individuals from the 10 existing populations, they discovered that the species is affected by inbreeding and a significant load of deleterious mutations. These advancements offer greater opportunities for the study of evolution and the conservation of plant resources.

In fact, the natural habitat and population of *C. pilosula* are continuously shrinking, suffering from the worsening climate change and indiscriminate harvesting, which is threatening the sustainable development and genetic diversity of *C. pilosula*. Therefore, generating a high-quality, chromosome-level genome for *C. pilosula* is essential for protecting genetic diversity, understanding the metabolism of its bioactive compound, and providing valuable insights into the evolutionary biology of this significant lineage.

In this study, by combining the PacBio sequencing and Hi-C technology, this effort resulted in a chromosome-level genome assembly of *C. pilosula*. Furthermore, we conducted a comparative genomic analysis on *C. pilosula* with 13 others species. The polyploidization histories of species within the Platycodonoid clade were validated through combined Ks and syntenic depth analyses. The results showed that duplicated genes, especially those resulting from TD and WGD, were the primary factors responsible for gene family expansion. The insights from this study are invaluable for the future conservation and sustainable utilization of this horticulturally and medicinally important plant species.

Materials and methods

Plant material and sequencing

Individuals of *C. pilosula* cultivated in the greenhouse of Yunnan Agricultural University were used for sequencing. Fresh leaves were harvested, immediately stored in liquid nitrogen, and subsequently sent to Novogene Bioinformatics Technology Co., Ltd. (Beijing, China) for sequencing. High-molecular-weight DNA was extracted using a modified CTAB method (Allen et al., 2006). The purity and concentration of the extracted DNA were evaluated using 1% agarose gel, and a Qubit fluorometer (Shanghai, China).

For short-read sequencing, a paired-end library was prepared using the NEBNext® Ultra™ Library Prep Kit and sequenced on the Illumina NovaSeq 6000 platform with a read length of 150 bp (Illumina, San Diego, CA, USA). For long-read sequencing, a Continuous Long Read (CLR) SMRTbell library was prepared using the SMRTbell Express Template Prep Kit 2.0 (Pacific Bioscience, CA, USA) according to the manufacturer's instructions. The long-read sequencing of *C. pilosula* DNA was then conducted on the PacBio Sequel platform (Pacific Bioscience, CA, USA).

A Hi-C library was constructed to generate chromosome-scale assembly. The fresh leaves were collected to construct the Hi-C libraries according to the previous library preparation protocol (Belton et al., 2012). Briefly, samples underwent vacuum infiltration cross-linking for 30 minutes with 3% formaldehyde at 4°C, followed by quenching with 0.375 M glycine for 5 minutes. After lysis of the cross-linked samples, endogenous nucleases were inactivated using 0.3% SDS. Chromatin DNA was then digested with 100 U MboI (New England Biolabs, Ipswich, MA, USA), labeled with biotin-14-dCTP of Invitrogen (Thermo Fisher Scientific, Waltham, MA, USA), and ligated using 50 U T4 DNA ligase (NEB, USA). Cross-links were reversed, and ligated DNA was purified using the QIAamp DNA Mini Kit (Qiagen, Hilden, Germany) following the manufacturer's instructions. The purified DNA was sheared into 300 to 500 bp fragments and was further blunt-end repaired, A-tailed and adaptor added, followed by purification through biotin-streptavidin–mediated pull-down and PCR amplification. Finally, the Hi-C libraries were quantified and sequenced on the Illumina NovaSeq 6000 platform (Illumina, San Diego, CA, USA) with a mode of paired-end 150 bp.

For protein coding gene annotation, RNA was separately extracted from three tissues (leaf, stem, and root) from the same *C. pilosula* individual using RNAprep pure Plant Kit (TIANGEN, China). RNA libraries were generated using NEBNext® Ultra™ RNA Library Prep Kit for Ilumina® (NEB, USA) following manufacturer's instruction and sequenced on Illumina NovaSeq 6000 platform. To obtain the transcripts, raw reads were trimmed using Fastp v0.20.1 (Chen et al., 2018) with the following parameters "-q 20 -l 70" and then assembled with Trinity v2.11.0 using the following parameters "–seqType fq –max_memory 100G –CPU 30" (Grabherr et al., 2011).

Genome assembly and annotation

The genome size was estimated using the Jellyfish v2.0 with the following parameters "-m 17 -s 100G -t 30 -c 7 -C" (Marçais and Kingsford, 2011) and GenomeScope v2.0 with default parameters (Ranallo-Benavidez et al., 2020). To achieve a high-quality genome assembly for *C. pilosula*, we utilized three primary long-read assembly software packages: NextDenovo v2.4.0 (https://github.com/Nextomics/NextDenovo) using the following parameters "task = all; rerun = 3; read_cutoff = 1 k; seed_cutoff = 8 k; seed_cutoff = 8k; genome_size = 683 Mb m; seed_cutfiles = 80; blocksize = 10 g; pa_correction = 80; minimap2_options_raw = -x ava-pb -t 16; sort_options = -m 10g -t 16 -k 50; correction_options = -p 32 random_round = 100 minimap2_options_cns = -x ava-ont -t 20 –k 17 –w 17; nextgraph_options = -a 1", MaSuRCA v4.1.0 using hybrid mode (Zimin et al., 2013), and Flye v2.9.3-b1797 using a parameter of "–threads 30" (Kolmogorov et al., 2019). The raw assembly was then polished by combining Pacbio CLR reads with Illumina short reads using NextPolish v1.3.1 with default parameters (Hu et al., 2020) for two rounds, followed by the removal of allelic contigs with Purge_Haplotigs v1.1.1 using default settings (Roach et al., 2018). The resulting contigs were scaffolded into chromosome-level scaffolds using Juicer v1.6.2 (Durand et al., 2016) and 3D-DNA pipeline (Dudchenko et al., 2017) with default parameters. To obtain the final genome assembly, the assembly errors (misjoins, misplacements, and orientation errors) in the scaffolds were manually corrected based on Hi-C contact signals using Juicebox v1.13.01 (https://github.com/aidenlab/Juicebox) (Robinson et al., 2018). The quality of the genome was evaluated using BUSCO v5.1.2 (Manni et al., 2021) with dataset embryophyta_odb10 (1,614 BUSCOs). Furthermore, the assembly quality value (QV) was assessed using Merqury v1.4 with a parameter of "k=19" (Rhie et al., 2020).

To identify repeat sequences, we utilized a combination of homology-based predictions and *de novo* predictions. Long Terminal Repeat (LTR) was identified using LTR_FINDER_parallel v1.1 (Ou and Jiang, 2019) with the following parameters "-harvest_out -size 1000000 -time 300 -finder" and LTRharvest v1.0 using the following parameters "-minlenltr 100 -maxlenltr 7000 -mintsd 4 -maxtsd 6 -motif TGCA -motifmis 1 -similar 85 -vic 10 -seed 20 -seqids yes" (Ellinghaus et al., 2008). Subsequently, the LTR candidates were filtered and the LTR Assembly Index (LAI) was calculated using LTR_retriever v2.8 (Ou and Jiang, 2018) with default parameters. Novel repetitive elements were predicted using RepeatModeler v2.0 with a parameter of "-engine rmblast" (Flynn et al., 2020). The predicted repeat libraries from LTR_retriever and RepeatModeler were combined and utilized by RepeatMasker v4.0.9 (http://www.repeatmasker.org) (Tarailo-Graovac and Chen, 2009) for *de novo* prediction with the following parameters "-a -nolow -no_is -norna". Repetitive elements were annotated with RepeatMasker and RepeatProteinMask v4.0.9 using the parameters of "-engine ncbi -noLowSimple -pvalue 0.0001" with Repbase v24.06 set as database (Bao et al., 2015). Tandem repeats were annotated with Tandem Repeat Finder v4.09 using the parameters of "2 7 7 80 10 50 2000 -d -h" (Benson, 1999). The results from these two annotations were combined to produce the final non-redundant repeat annotation. The chromosomal distributions of Ty3-retrotransposons and Ty1-retrotransposons in *C. pilosula* were calculated using a sliding window of 1 Mbp.

Protein-coding genes were predicted by combining *de novo*, homology-based, and transcript-based methods. Augustus v3.2.2 (Stanke et al., 2008) was used for the *de novo* gene prediction. For the homology-based method, protein sequences of *Arabidopsis thaliana* (L.) Heynh. (GCF_000001735.4) (Willing et al., 2015), *Citrus sinensis* (L.) Osbeck (GCF_022201045.2) (Wu et al., 2023), *Solanum lycopersicum* L. (GCF_000188115.5) (Sato et al., 2012), and *Vitis vinifera* L. (GCF_030704535.1) (Shi et al., 2023) were downloaded from the National Center for Biotechnology Information (NCBI) and aligned to the genome of *C. pilosula* using TBLASTN v2.2.29+ (Camacho et al., 2009) with an E-value threshold of $1e^{-5}$. GeneWise v2.4.1 (Birney et al., 2004) was then used to predict gene models based on these alignments with default parameters. For the transcripts-based method, the assembled transcripts were mapped to the genome and analyzed using Program to Assemble Spliced Alignments (PASA) v2.4.1 (Haas et al., 2008) to predict genes. Finally, a consensus gene model was combined by EvidenceModeler v1.1.1 (Haas et al., 2008) with the annotated gene structures from *de novo*, homology-based, and transcripts-based methods. Additionally, alternatively spliced sites and untranslated regions (UTRs) were incorporated using PASA v2.4.1 (Haas et al., 2008). For functional annotation of protein-coding genes, we employed eight public databases, including Uniprot, TrEMBL, GenBank NR, KEGG, EggNOG, GO, InterProScan, and Pfam. The annotation process was performed using DIAMOND v0.9.14.115 (Buchfink et al., 2021) with a threshold of e-value ≤ $1e^{-5}$. The identification of transfer RNAs (tRNAs), was performed using tRNAscan-SE v2.0.7 (Chan et al., 2021). Other non-coding RNAs (ncRNAs), such as microRNAs (miRNAs), ribosomal RNAs (rRNAs), and small nuclear RNAs (snRNAs), were identified using Infernal v.1.1 (Nawrocki and Eddy, 2013) by searching against the Rfam v.14.1 (Kalvari et al., 2021) database.

Comparative genomics and phylogenetic analyses

The longest protein sequences from *C. pilosula*, along with 13 other species (Supplementary Table S3), were clustered into protein

sequence in these groups using OrthoFinder v2.5.2 with parameters "-t 30 -a 20 -M msa" (Emms and Kelly, 2019). Single-copy orthologous groups shared by all species were identified and, each protein sequence in these groups was individually aligned using MAFFT v7.475 with parameters "–localpair –maxiterate 1000" (Katoh and Standley, 2013). Subsequently, the corresponding coding sequences (CDS) were aligned to the codon alignments according to the alignments of these protein-coding sequences using PAL2NAL v14 (Suyama et al., 2006). Poorly aligned regions within these codon alignments were filtered out using trimAl v1.4.rev15 with a parameter of "-automated1" (Capella-Gutiérrez et al., 2009). Finally, those codon alignments of the single-copy orthologous groups were concatenated to build a Maximum Likelihood (ML) phylogenetic tree using IQ-TREE v2.2.0.3 with the parameters "-m MFP -bb 1000 -nt 10" and the best-fit model (GTR + F + I + G4) (Nguyen et al., 2015). Divergence times were estimated based on ML tree using MCMCTree v4.10.0 (dos Reis, 2022) from the PAML (Yang, 2007) package with parameters of "burnin=50000; nsample=100000". Two calibration points were sourced from the TimeTree database (http://www.timetree.org/). The first calibration involved a comparison between *Oryza sativa* and *V. vinifera*, dated between 142 to 163 million years ago (Mya). The second calibration compared *B. vulgaris* with *V. vinifera*, with an estimated divergence time ranging from 111 to 124 Mya. The phylogenetic tree, including the divergence times, was visualized using FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/). The gain and loss of orthogroups among 14 species along the phylogenetic tree were estimated using CAFE v5 (Mendes et al., 2021) with a *P* value threshold of 0.05. The GO enrichment analysis of expansion genes was conducted via clusterProfiler v4.2.2 with a *P* value threshold of 0.05 (Wu et al., 2021).

Synteny and polyploidization exploration

The intraspecific synteny was analyzed using the WGDI toolkit v0.5.1 (Sun et al., 2022). First, intraspecific homologs were extracted using BLASTP v2.2.29+ (Camacho et al., 2009) with an e-value cutoff of $1e^{-5}$. Collinear gene pairs were identified by WGDI (Sun et al., 2022) with a parameter of '-icl'. Ks values of collinear gene pairs were calculated using parameter '-ks' in WGDI with Nei-Gojobori method (Nei and Gojobori, 1986). The medium Ks values of collinear blocks were fitted through Gaussian kernel density estimation by WGDI (Sun et al., 2022) with a parameter of '-pf' and plotted using a parameter of '-kf'. The time points of WGD events were calculated according to $T = Ks/2r$, where r represents a substitution rate of 6.5×10^{-9} mutations per site per year for eudicots (Tu et al., 2020). The interspecific synteny patterns of *V. vinifera vs. C. lanceolata*, *C. lanceolata vs. C. pilosula*, and *C. pilosula vs. P. grandiflorus* were also investigated by JCVI (Tang et al., 2024). The dot plot between *C. pilosula vs. P. grandiflorus* was visualized by WGDI (Sun et al., 2022) with '-d' parameter (Sun et al., 2022). Based on the characterized ancestral karyotype of core eudicots and the inferred polyploidization history (Sun et al., 2022), the common ancestral karyotype between *C. pilosula* and *P. grandiflorus* following the γ-WGT event was illustrated using Adobe Animate software and WGDI (Sun et al., 2022).

The identification of different modes of gene duplication and the analysis of CYP superfamily

Various gene duplication modes were identified utilizing the "DupGen_finder-unique.pl" module of DupGen_finder (Qiao et al., 2019) with default parameters, and *B. vulgaris* was set as the reference. We identified cytochrome P450 (CYP) and UDP-Glycosyltransferase (UGT) gene family using HMMER v3.1b2 (Potter et al., 2018) with parameter 'E-value $1e^{-5}$'. The Pfam HMM models, namely PF00067 and PF00201, were set as queries for the identification of CYP and UGT genes, respectively. The previously characterized *A. thaliana* CYP and UGT genes were downloaded from the database of "The Arabidopsis Cytochrome P450, Cytochrome b5, P450 Reductase, β-Glucosidase, and Glycosyltransferase Site" (http://p450.kvl.dk/index.shtml) and used as outgroups. To construct the phylogenies for CYP genes, the protein sequences were aligned using MAFFT v7.475 (Katoh and Standley, 2013) followed by trimming with trimAl v1.4.rev15 (Capella-Gutiérrez et al., 2009). ML phylogenetic trees were constructed with IQ-TREE v2.2.0.3 (Kalyaanamoorthy et al., 2017) and visualized with the online tool iTOL (https://itol.embl.de/) (Letunic and Bork, 2024).

Results

Genome sequencing, assembly, and annotation of *C. pilosula*

For the assembly of *C. pilosula* genome, we employed long-read PacBio SMRT sequencing, complemented with short-read Illumina NovaSeq 6000 sequencing for error correction. Before assembling the genome, a survey was performed using 80 Gb of Illumina short reads with a K-mer size of 21 (Supplementary Table S1). This analysis determined the genome size to be 683.47 Mb and the heterozygosity rate to be 0.96% (Figure 1A). In total, 102.77 Gb of PacBio reads (~150 fold genome coverage) were generated for the *de novo* assembly of *C. pilosula* genome (Supplementary Table S1). The assembly process using the NextDenovo v2.4.0 generated about 855.06 Mb of sequences with a contig N50 of 2.05 Mb, the MaSuRCA (Hybrid Illumina and PacBio assembly) produced around 712.45 Mb of sequences with a contig N50 of 0.34 Mb, and the Flye generated about 933.67 Mb of sequences with a contig N50 of 0.19 Mb (Supplementary Table S2). The BUSCO completeness evaluation revealed that the assembly completeness of all three methods exceeds 97% (Supplementary Table S2). Furthermore, based on the results, NextDenovo assembly displayed superior contiguity than other methods, so its primary assembly was selected for further analysis. After two rounds of polishing and the removal of allelic contigs, the resulting draft assembly of *C. pilosula* had a total length of 679.20 Mb and a contig N50 of 2.28 Mb (Supplementary Table S2). Next, 96.5% of the draft assemblies were anchored onto eight pseudo-chromosomes in the chromosome-level genome assembly with the 224.89 Gb Hi-C data (~329.26 fold genome coverage) (Supplementary Table S1),

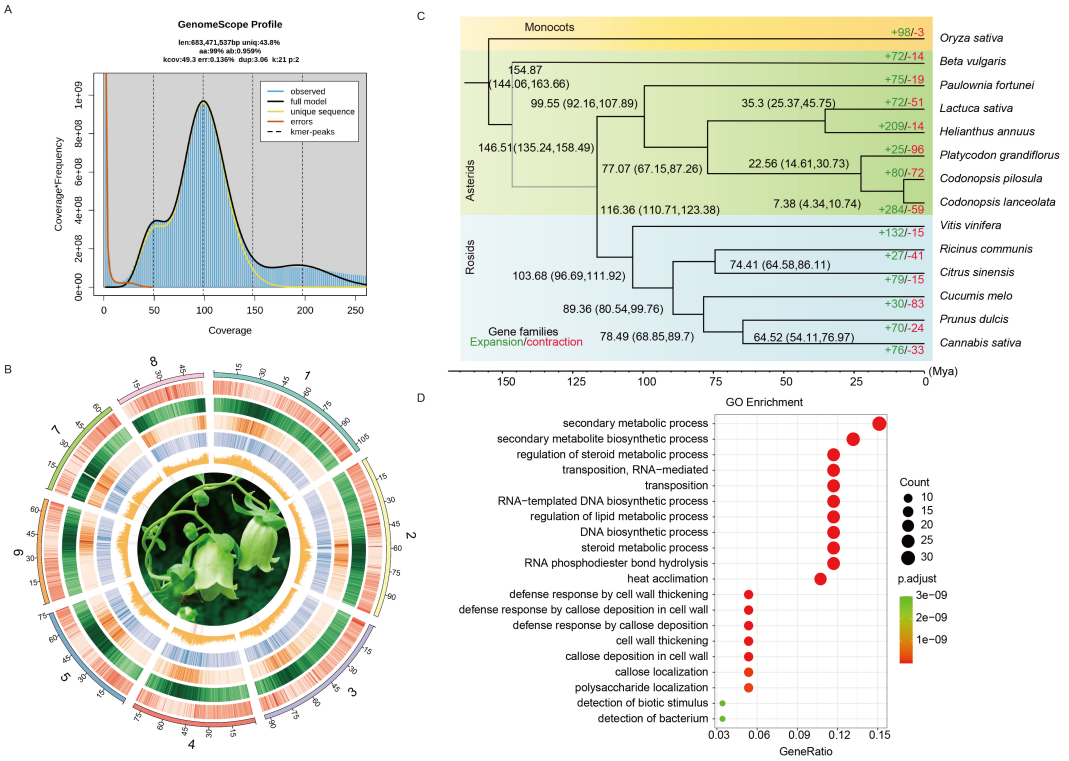

FIGURE 1
Genomic features and comparative analysis of *C. pilosula* with 13 other species. **(A)** Genome survey of *C. pilosula*. **(B)** The genomic features are arranged in the order of pseudo-chromosomes (scale is in Mb), gene density, repeat density, Ty3-retrotransposons, Ty1-retrotransposons, and GC contents, and flower photos from outside to inside in 1 Mb intervals across the 8 pseudo-chromosomes. **(C)** Phylogenetic tree inferred from single-copy orthologs among selected species. Black numbers in each node denoted the divergence time of each clade (Mya). The green and the red numbers at the terminal branches showed the expansion (green) and contraction (red) of gene families for each species. **(D)** GO enrichment analysis of rapidly expanded genes for *C. pilosula*.

resulting in a total assembly length of 679.60 Mb and a scaffold N50 of 80.50 Mb (Figure 1B; Supplementary Figure S1; Supplementary Table S2). The size of the final assembled genome was close to the estimations (Figure 1A). BUSCO completeness analysis indicated that the final genome assembly contained approximately 97.6% of the embryophyta_odb10 BUSCO gene sets, which is comparable to those of related species (Supplementary Table S3). Overall, all Illumina reads were mapped to the *C. pilosula* assembly, achieving a mapping rate of 99.13% and a genome coverage rate of 96.14% (Supplementary Table S4). Merqury evaluation revealed that the genome was 91.64% complete with a QV value of 36.50.

For the repetitive elements annotation, the combination of homology-based and *ab initio* predictions revealed that the *C. pilosula* genome contains 70.26% non-redundant repetitive elements. Among these, LTR were the most abundant transposable elements (TEs), constituting 39.17% of the total, while DNA transposons trailed at 8.02% and long interspersed nuclear elements (LINEs) made up 3.77% (Supplementary Table S5). Evaluation of the LAI revealed that the genome assembly had an LAI of 14.89, which meets the reference grade suggested by Ou et al. (2018) (Supplementary Table S5).

By combining transcriptome-, homology-, and *ab initio*-based techniques, a consensus gene model was generated, resulting in the prediction of 29,808 protein-coding genes, representing 96.50% of the embryophyta_odb10 BUSCO gene sets, which could be comparable with the relative species (Supplementary Table S6). Functional annotation indicated that 98.98% of the *C. pilosula* genes could be annotated in at least one of the existing databases, including SwissProt, NR, TrEMBL, KEGG, EggNOG, GO, Pfam, and Interproscan (Supplementary Table S7). Furthermore, we identified 912 tRNAs, 345 rRNAs, 137 miRNAs, and 598 snRNAs (Supplementary Table S8).

Phylogenetic and comparative genomics analyses

Based on the clustered results of OrthoFinder v2.5.2, a Maximum Likelihood (ML) phylogenetic tree was constructed using 1,004 single-copy genes present in all 14 species (Supplementary Table S9). The analysis revealed that *C. pilosula* was the most closely related to *C. lanceolata* and is also clustered with *P. grandiflorus*. According to the time-calibrated molecular clock, we estimated that Campanulaceae diverged from other Asterids approximately 67.15–87.26 million years ago (Mya), while *C. pilosula* diverged from *P. grandiflorus* approximately 14.61–30.73 Mya (Figure 1C).

Among the 29,808 protein-coding genes identified in the *C. pilosula* genome, 27,856 genes were grouped into 15,574 families (Supplementary Table S10). Within these families, 773 were unique to *C. pilosula*, while 8,853 families were shared with the other 13 studied species (Supplementary Tables S9, S10). GO enrichment analysis revealed that these unique paralogous genes were highly enriched in the biological process (BP) terms "DNA replication", "trichoblast maturation", and "root hair cell differentiation"; the cellular component (CC) termed "extracellular space" and "nuclear body"; and the molecular function (MF) termed "identical protein binding" and "aminoacylase activity" (Supplementary Table S11). Interestingly, some of these genes were closely associated with the development of root hairs.

Gene contraction and expansion analysis showed that, since diverging from *P. grandiflorus*, the *C. pilosula* genome had experienced more gene family expansions (80) than contractions (72) (Figure 1C). GO enrichment analysis revealed that the expanded paralogous genes were most enriched in the biological process (BP) terms "secondary metabolic process", "secondary metabolite biosynthetic process", and "regulation of steroid metabolic process"; the cellular component (CC) terms "retrotransposon nucleocapsid" and "SCF ubiquitin ligase complex"; and the molecular function (MF) terms "RNA-directed DNA polymerase activity" and "DNA-directed DNA polymerase activity". Interestingly, these genes are closely associated with steroid and lipid metabolism (Figure 1D; Supplementary Table S12). In the contracted paralogous gene sets, the GO enrichment analysis revealed that more genes enriched in the biological process (BP) terms "response to ethanol"; the cellular component (CC) terms "lysosome"; and the molecular function (MF) terms "beta-glucosidase activity" (Supplementary Table S13).

The polyploidization and karyotype evolutionary history of platycodonoids clade species

To investigate the ancient polyploidization history of platycodonoid clade species in the chromosome-level genome, we analyzed the distribution of substitutions per synonymous site (Ks) in intra-genomic collinear blocks. The Ks analyses revealed that *C. pilosula*, *C. lanceolata*, and *P. grandiflorus* exhibited a single Ks peak, similar to *V. vinifera*, indicating that they all underwent the γ-WGT event shared by all core eudicot species (Figure 2A). The distribution of Ks peaks showed a wider range, from 1.14-1.76 (120-135.38 Mya), the lowest Ks value of 1.14 (87.69 Mya) in *V. vinifera* (Figure 2A). Among these three platycodonoid clade species, *C. pilosula* exhibited the lowest Ks value of 1.68 (129.23 Mya), representing the slowest evolutionary rate. However, this Ks distribution pattern conflicted with the previous results reported by Crowl et al. (2016), which indicated that *P. grandiflorus* had two Ks peaks at 0.5 and 1.55, representing two WGD events. To further validate this polyploidization history, the analysis of interspecies synteny relationships was conducted. Interspecies synteny comparisons between *V. vinifera* and the three Campanulaceae species suggested that for each *V. vinifera* genomic region there was up to one syntenic region in *C. pilosula*, *C. lanceolata* and *P. grandiflorus*. Syntenic depth comparison of *C. pilosula vs. V. vinifera*, *P. grandiflorus vs. V. vinifera*, and *C. pilosula vs. P. grandiflorus* all exhibited a consistent 1:1 pattern (Figure 2B). These analyses confirmed that no additional WGD occurred in these three Campanulaceae species following the γ-WGT.

Synteny analysis revealed that the genomes of *C. pilosula* (x=8) and *C. lanceolata* (x=8) shared the same chromosome numbers and displayed a 1:1 synteny relationship, along with no extra chromosome fusions or fissions (Figure 2B). However, the chromosome numbers differed slightly between *C. pilosula* (x=8) and *P. grandiflorus* (x=9). Synteny analysis revealed that six chromosomes in both species might have originated directly from six ancestral chromosomes through several inversions (Figure 2B). In contrast, the formation of *C. pilosula* chromosomes 1 and 2 and *P. grandiflorus* chromosomes 2, 4 and 8 was more complicated. Previous studies indicated that chromosomal evolution in land plants was mainly characterized by descending dysploidy, including nested chromosome fusion (NCF) or end-to-end joining (EEJ), as well as non-dysploid changes like inversions, reciprocal translocations (RT), deletions, and duplications (Sun et al., 2022). Based on these findings, we speculated that ancestral chromosomal fusion and non-dysploid changes contributed to the current chromosome structure of *C. pilosula* and *P. grandiflorus*. To infer the likely evolutionary trajectories underlying these chromosomal changes, we visualized the synteny dot plot between them. The *P. grandiflorus* chromosome 8 represented an intact structure of telomere to telomere. Therefore, this intact chromosome was tacked as an ancestral chromosome according to the 'Telomere-centric genome repatterning model' proposed in the previous study (Sun et al., 2022) (Figure 2C). Then, the synteny relationship between *C. pilosula* chromosomes 1 and 2, and *P. grandiflorus* chromosomes 2 and 4, could be explained by the reciprocally translocated chromosome arms (RTA) (https://github.com/SunPengChuan/wgdi-example/blob/main/Karyotype_Evolution.md) (Sun et al., 2022). However, the history of this RTA is difficult to distinguish according to only these two species. Thus, the interspecies synteny relationships between them and *V. vinifera* were visualized, respectively. Because the *V. vinifera* genome represents the most ancestral eudicot karyotype, it served as the reference (Sun et al., 2022).

Based on the interspecies synteny relationships, the homologous synteny blocks between *V. vinifera* chromosomes 1 and 18, and *P. grandiflorus* chromosomes 2 and 4, were disrupted, whereas this pattern was absent between *V. vinifera* and *C. pilosula* chromosomes 1 and 2 (Supplementary Figures S2, S3). Furthermore, *V. vinifera* chromosomes 18 was previously proved to represent an intact ancestral chromosome of the eudicot karyotype (Sun et al., 2022). To further verify if the results were due to assembly errors, a collinear analysis was performed using data from Jia et al. (2022) and Lee et al. (2023). This analysis showed that homologous chromosomes, especially chromosomes 2 and 4, had stronger collinear relationships, indicating no significant assembly errors (Supplementary Figure S4). Therefore, according

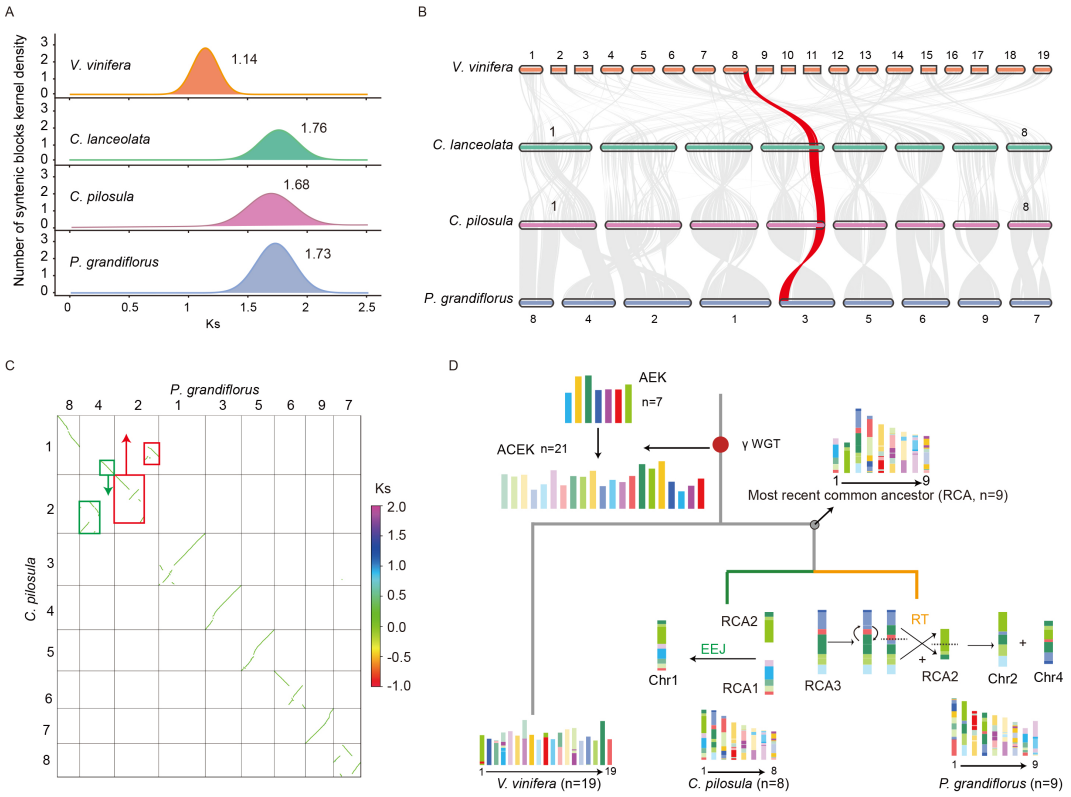

FIGURE 2
Polyploidization history and karyotype evolution in three platycodonoids species. **(A)** Distribution of the numbers of synonymous substitutions per synonymous site (Ks) of three platycodonoids species and *V. vinifera*. **(B)** The collinearity relationships of three platycodonoids species and *V. vinifera*. **(C)** The green and red boxes represented homologous synteny blocks that underwent reciprocal translocations (RT) in chromosomes 2 and 4 of *P. grandiflorus*. **(D)** Karyotype evolution of *C. pilosula* and *P. grandiflorus*.

to the 'Telomere-centric genome repatterning model', we believed that the *P. grandiflorus* chromosomes 2 and 4 experienced a RT, rather than the *C. pilosula* chromosomes 1 and 2. Based on those findings, we inferred that platycodonoids clade species had a common ancestral karyotype n=9 (Figure 2D).

Repetitive elements driving the evolution of the genome

Generally, the genome size variation could be largely attributed to the difference of the transposable elements (TEs) or the polyploidization events (Kapusta et al., 2017; Wang et al., 2021). *C. pilosula*, *C. lanceolata*, and *P. grandiflorus* shared a common polyploidization history according to the previous results in this study. However, the genome size of *C. pilosula* (679.60 Mb) is 9.05% larger than that of *P. grandiflorus* (622.86 Mb) and appreciably smaller by 46.62% compared to *C. lanceolata* (1,273.24 Mb) (Supplementary Table S3). Therefore, to reveal the causes for the variation in genome size among these three species, comparative analyses of their TEs evolution and contents were conducted. This analysis indicated that *C. lanceolata* had the highest TE content of 950.72 Mb (74.67%) among these three species. (Figure 3A; Supplementary Table S14). The *C. pilosula* and *P. grandiflorus* showed similar TE contents. LTR elements were the most abundant types among the classified TEs, constituting over 28% of the three species (Figure 3A; Supplementary Table S14).

To gain a clear understanding of the landscape of TEs in these three species, the chromosomal distribution of LTR elements, specifically Ty3-retrotransposons and Ty1-retrotransposons, was visualized. The abundance of Ty3-retrotransposons was higher than that of Ty1-retrotransposons in these three species. A comparable content of TEs was observed in the genomes of *C. pilosula* and *P. grandiflorus*, consistent with their TE contents (Figure 3B). The Ty3-retrotransposons and Ty1-retrotransposons were much more abundant in the pericentromeric regions compared to the chromosomal terminal regions in both species. However, this high density TE distribution in pericentromeric regions was not evident in the *C. lanceolata* genome (Figure 3B). This pattern was uncommon in the genomes of most reported species. Meanwhile, both Ty3-retrotransposons and Ty1-retrotransposons in *C. lanceolata* showed higher abundance compared to *C. pilosula* and *P. grandiflorus* (Figure 3B). Therefore, this higher TE contents were responsible for the expansion of genome size indeed.

We further estimated the insertion times of LTR in these three species. The insertions of the majority of the Ty1-retrotransposons started at ~0-1.5 Mya for the three species. The proliferation of Ty1-retrotransposons in *C. pilosula* and *P. grandiflorus* exhibited a

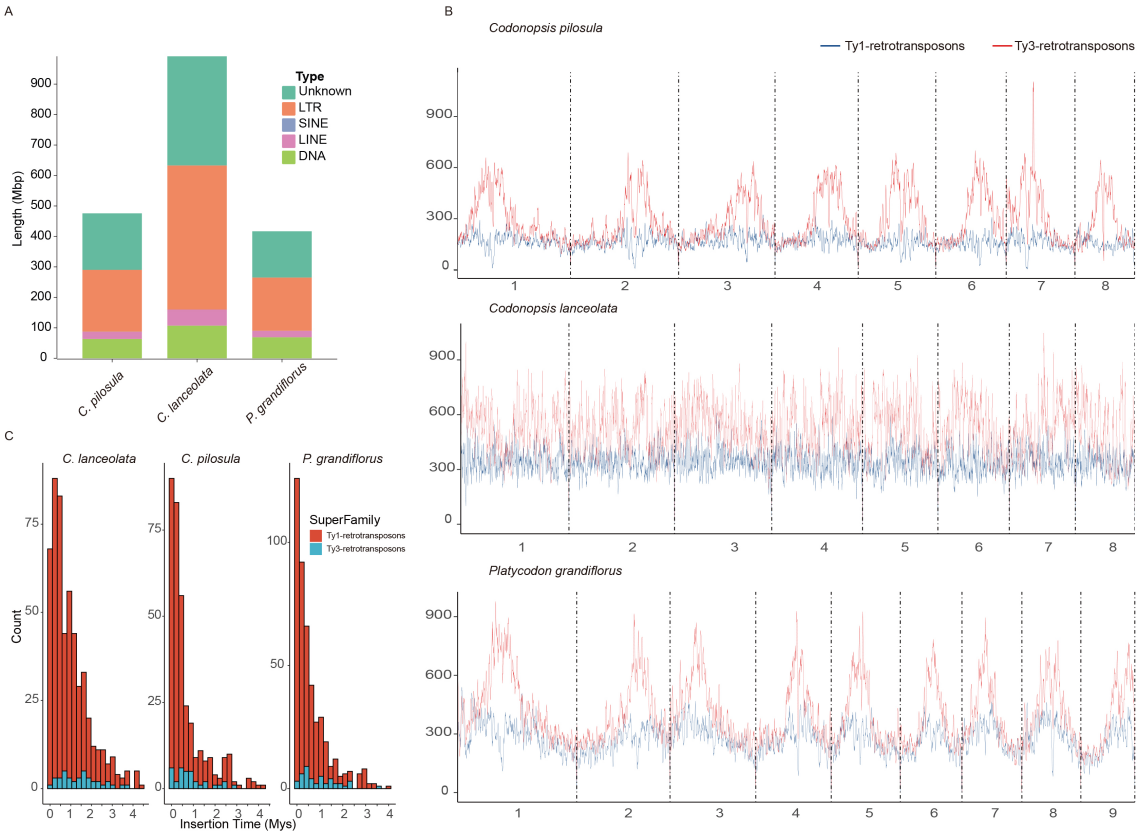

FIGURE 3
Evolutionary analysis of TEs in three platycodonoids species. **(A)** Sequence size statistics for various types of transposable elements (TEs). **(B)** LTR density distribution. The Ty1-retrotransposons indicated by a blue line and the Ty3-retrotransposons indicated by a red line. **(C)** Insertion time distribution of intact Long Terminus Repeat (LTR).

similar pattern, peaking around 0.1 Mya, slightly earlier than *C. lanceolata* at ~0.25 Mya. It is noted that *C. pilosula* and *P. grandiflorus* had an obvious Ty1-retrotransposons proliferation at ~2.5 and ~2.9 Mya, respectively. Moreover, the majority of Ty3-retrotransposons insertions at 0.5 Mya in *C. pilosula* and *P. grandiflorus*, which lagged behind the peaks of Ty1-retrotransposons. In *C. lanceolata*, the major insertions of Ty3-retrotransposons LTR occurred around at 0.5-2Mya. This higher density of insertions may explain its greater Ty3-retrotransposons content compared to the other two species in platycodonoid clades. *C. pilosula* also showed a minor increase in Ty3-retrotransposons proliferation around ~2.25 Mya, while this phenomenon was absent in the other two species (Figure 3C).

Duplicated genes driving the expansion of gene families

WGD events and duplicated genes are the primary driving force for the gene family expansion (Magadum et al., 2013). Generally, duplicated genes could be classified into five types including whole-genome duplication (WGD), tandem duplication (TD), proximal duplication (PD), dispersed duplication (DSD), and transposed duplication (TRD) (Qiao et al., 2019). Genes with the same Pfam domain are typically grouped into the same gene family. Therefore, to investigate the influence of duplicated genes onto the expansion of gene families, the genes with the Pfam annotation were classified into different duplicated gene types across the analyzed 13 species. This analysis showed that over 72% of the Pfam annotated genes could be classified into five duplicated gene types among these 13 species (Figure 4A; Supplementary Table S15). This higher proportion suggested that duplicated genes indeed played important roles in driving the expansion of gene families. Interestingly, *Helianthus annuus*, *Lactuca sativa*, and *Paulownia fortunei* showed a higher proportion of WGD genes compared to other species, due to their history of additional WGD events, with the exception of the γ-WGT event (Figure 4A). The Ks analyses of duplicated genes in *C. pilosula* showed that WGD and TRD genes exhibited a higher average Ks value distribution, resembling the Ks peak value of the γ-WGT event (Figures 2A, 4B). The Ks average value of TD, PD, and DSD genes were all less than 1, indicating that most of them could have been generated following the γ-WGT event (Figure 4B).

To illustrate the impact of duplicated genes in more details, two major gene families, CYP and UGT, were selected for the study. The distribution of duplicated genes in these two gene families notably differed from that of the total genes across the 14 species. The TD genes showed the highest level of abundance, accounting for almost

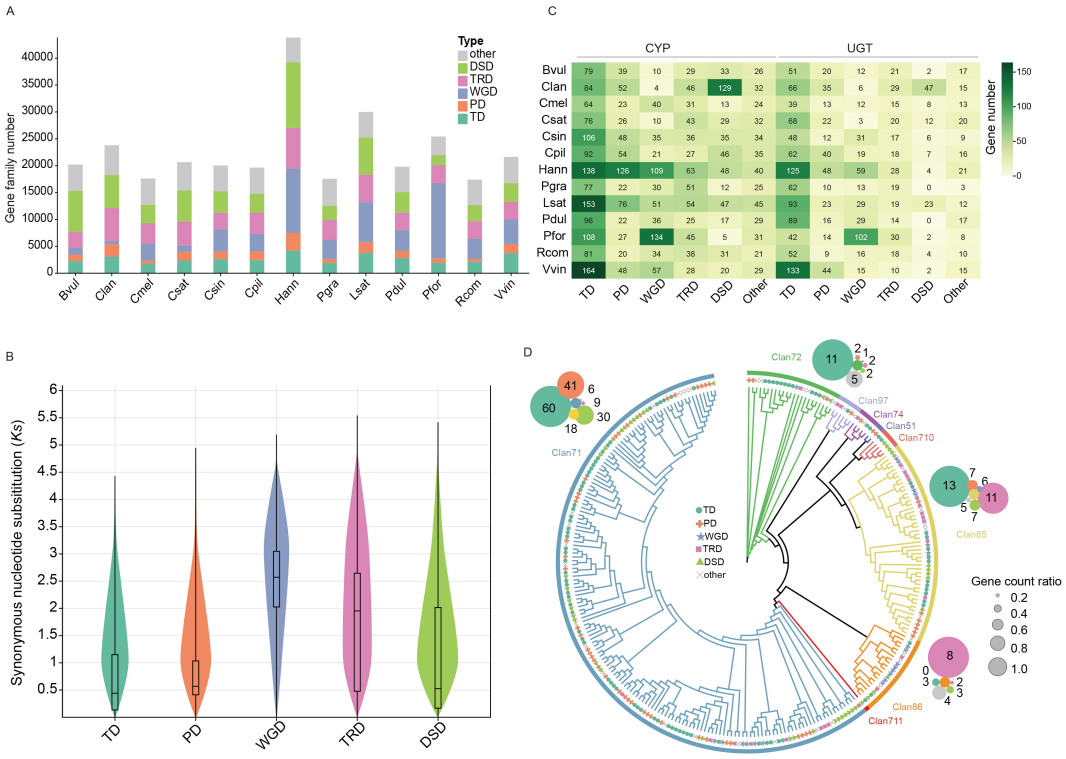

FIGURE 4
The numbers and distributions of various duplicated genes across 13 species. **(A)** The distribution of transposed duplicates (TRD), dispersed duplicates (DSD), tandem duplicates (TD), proximal duplicates (PD), whole-genome duplicate (WGD) and other genes. Bvul (*B. vulgaris*), Clan (*C. lanceolata*), Cmel (*C. melo*), Csat (*C. sativa*), Csin (*Citrus sinensis*), Cpil (*C. pilosula*), Hann (*Helianthus annuus*), Pgra (*Platycodon grandiflorus*), Lsat (*Lactuca sativa*), Pdul (*Prunus dulcis*), Pfor (*Paulownia fortune*), Rcom (*Ricinus communis*), Vvin (*Vitis vinifera*). **(B)** The distribution of synonymous substitutions per synonymous site (Ks) in *C. pilosula* for various modes of gene duplications. **(C)** The proportions of cytochrome P450 (CYP) and UDP-Glycosyltransferase (UGT) genes from different gene duplications. **(D)** The phylogenetic tree and the proportions of 9 Clan CYP genes from *C. pilosula*. The outer circles represent the counts of gene duplication modes for the four largest Clan CYP genes: Clan 71, Clan 85, Clan 72, and Clan 86.

1/3 of the CYP or UGT gene families across the 13 species, respectively (Figure 4C). It is worth noting that the PD and DSD genes also had a significant proportion in the *Helianthus annuus* and *C. lanceolata* CYP genes. Interestingly, although 10 out of the 13 species shared a common polyploidization history with the γ-WGT event, they exhibited a relatively low proportion of WGD genes in their CYP and UGT gene families. In contrast, the other three species, *H. annuus*, *L. sativa*, and *P. fortune*, which have undergone additional recent WGD events, showed a higher proportion of WGD genes in these gene families (Figure 4C). This phenomenon may be attributed to the loss or neofunctionalization of the majority of duplicated genes generated from γ-WGT event during the long evolutionary history. To investigate the duplicated genes distribution of CYP gene family in different clans, the phylogenetic tree of *C. pilosula* CYP genes was constructed. Totally, 275 CYP gene sequences were further clustered into nine clans. Clan711 is single-family CYP clan. Clan71 is the largest CYP clan, followed by Clan85, Clan72, Clan86, Clan97, Clan710, Clan74, and Clan51. Clan71 comprised over half of all genes and covered all types of duplicated genes (Figure 4D). Among them, four larger clans, namely Clan71, Clan85, Clan72, and Clan86, were selected to investigate the distribution of duplicated genes. The analysis of the duplicated genes showed that over 1/3 of Clan71 CYP genes originated from TD genes, followed by PD, DSD, other, TRD, and WGD. Within Clan85 and Clan72, TD genes retained their position as the most prevalent. Conversely, Clan86 showcased the highest proportion of WGD genes (Figure 4D).

Discussion

The completion of the chromosome-level genome assembly for *C. pilosula* marks a significant advancement in deciphering the genomic intricacies of this species. The integration of multiple sequencing technologies and various assembly strategies has resulted in a comprehensive genomic resource. This lays the groundwork for in-depth studies of the *C. pilosula* genetic landscape. Rigorous validation steps, such as mapping Illumina short reads, LAI index, and BUSCO evaluations, have collectively affirmed the reliability, accuracy, and completeness of the *C. pilosula* genome assembly (Supplementary Tables S2, S5). These metrics underscore the technical robustness of the assembly process and provide confidence for subsequent analyses and interpretations.

The complete *C. pilosula* genome assembly enhances our understanding of both the coding and non-coding components of the genome, which is important for future studies on gene regulation and functional genomics.

By integrating the annotated *C. pilosula* genome with previously published genomes representing Campanulaceae members, we identified one round of WGD event (γ-WGT) in *C. pilosula*, *C. lanceolata*, and *P. grandiflorus* (Figure 2A). The observed collinearity between the genome, along with comparisons to other species, strongly supported this γ-WGT event (Figures 2B, C). However, this finding contradicts the previous two rounds of WGD events proposed by Crowl et al. (2016), which were only inferred from transcriptome data. Generally, WGD inference based on Ks values was limited due to the difficulty in distinguishing gene pairs originating from WGD and those arising from small-scale duplications (SSDs) without structure information (Zwaenepoel et al., 2018). For example, some misinterpretations of the correspondence between Ks values and WGDs based on transcriptome alone had been demonstrated in some species, such as *Callicarpa americana* (Hamilton et al., 2020), watermelon (Guo et al., 2013), black pepper (Hu et al., 2019), Olive (Ren et al., 2018), and *Prunus mongolica* (Zhu et al., 2023). In summary, integrating genomic collinearity analysis with Ks information provided a more accurate and effective method for inferring polyploidization events, as supported by our findings in this study and previous research (Kong et al., 2023; Sun et al., 2024).

Following the WGD events, numerous species underwent the post-polyploid diploidization (PPD) processes, as evidenced in various specific lineages (Mandáková and Lysak, 2018). PPD processes mainly involve changes in genome size, chromosomal rearrangements, subgenome-specific fractionation, including biased gene retention/loss and gene sub-/neofunctionalization, and others (Cheng et al., 2018; Mandáková and Lysak, 2018). Among them, chromosomal rearrangements, including EEJ, NNC and RT (Sun et al., 2022), represented the most dramatic type of karyotype evolution, injecting significant potential for the species diversity and speciation. According to the 'Telomere-centric genome repatterning model' proposed in the previous study (Sun et al., 2022), the platycodonoids clade species had a common ancestral karyotype n=9 (Figure 2D). Subsequently, *C. pilosula* and *C. lanceolata* underwent a single EEJ fusion to form the present karyotype n=8; *P. grandiflorus* experienced a single RT without any additional chromosome fusion events to achieve the current karyotype n=9 (Figure 2D). This *Codonopsis*-specific EEJ fusion could play a crucial role in driving the speciation of *Codonopsis* and *Platycodon*.

In the plant kingdom, species vary in genome size from a few hundred Mbs to tens of gigabytes (Pellicer et al., 2018). This variation is typically attributed to factors such as polyploidization, gene duplications, repeat expansion, and other events (Wendel et al., 2016; Wang et al., 2021; Zuntini et al., 2024). In our study, the findings indicated that three platycodonoid clade species has only undergone the γ-WGT event. The shared polyploidization history of these species provides ideal materials for a deeper exploration of the mechanisms underlying their unique genomic expansion. Our analyses of repetitive elements indicated these repetitive elements play an important role in driving genome expansion and causing variations in genome size. The significance of repetitive elements in driving genome size has also been confirmed in other species, such as hawthorn (Zhang et al., 2022), *Welwitschia* (Wan et al., 2021), and *Cycas* (Liu et al., 2022). LTR elements account for almost half of the repetitive elements. The analyses of insertion times indicated that the proliferation of LTR elements varied among these three species. Interestingly, many LTR elements burst took place after their speciation, suggesting that this differing proliferation could be derived from their adaptive evolution. However, this adaptive assumption is awaiting more conclusive evidence.

Polyploidization and gene duplication have been important driving forces of genome evolution, playing a crucial role in adaptation to new environments (Liu et al., 2023; Ye et al., 2024). Our analyses of different duplication modes in gene families revealed that WGD, TD, PD, DSD, and TRD have profound implications for the expansion of gene families, with TD exhibiting the most obvious impact. WGD also apparently influenced the gene families of *H. annuus*, *L. sativa*, and *P. fortune* as they all underwent additional WGD events following the γ-WGT event. In contrast, this phenomenon is less apparent in the other 11 species, which only experienced the γ-WGT event, possibly due to the drastic loss of most WGD genes over their long evolutionary history (Gout et al., 2023). Furthermore, two specific gene families, CYP and UGT, were included in this analysis. The results showed that many genes tend to be clustered within TD, further indicating its important contributions to gene family evolution. In summary, polyploidization and other duplicated genes play important roles in driving the evolution of gene families, commonly expanding them. However, the dynamic influence on gene families, as well as the corresponding gene fate, remain understudied topics.

In conclusion, this study presents the first high-quality haploid genome assembly for *C. pilosula*. This marks the first step towards understanding the molecular basis of various desirable traits in economically important species. Although more research is needed on chromosomal structural diversity and haplotype-resolved genomes, our findings provide a basis for future research in comparative genomics, molecular biology, genetics, and evolutionary aspects of species of the platycodonoid clade.

Data availability statement

The raw genome sequencing data of *C. pilosula* are available at the National Center for Biotechnology Information (NCBI) under BioProject number PRJNA1068481. The raw sequence and genome sequence (https://ngdc.cncb.ac.cn/gwh/Assembly/85966/show, GWHFAJJ00000000.1) have also been uploaded to NGDC database under the accession of PRJCA029124 (CRA018369). Besides, the annotation results, including protein coding genes and repeat annotation, have been uploaded to the Figshare database (https://doi.org/10.6084/m9.figshare.26799025). All data are available from the corresponding author upon request.

Author contributions

B-ZC: Writing – original draft, Writing – review & editing, Data curation, Methodology, Validation, Visualization. Z-JY: Visualization, Writing – review & editing. LY: Writing – review & editing, Resources. Y-FZ: Writing – review & editing. X-ZL: Writing – review & editing. LW: Writing – review & editing. Y-PZ: Writing – review & editing. G-HZ: Writing – review & editing. D-WL: Validation, Writing – review & editing. YD: Conceptualization, Validation, Writing – review & editing. S-CD: Conceptualization, Validation, Visualization, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was supported by National Key R&D Program of China from the Ministry of Science and Technology of China (grant No. 2019YFC1711100) and Digitalization and Utilization of Biological Resources (grant No. 202002AA100007).

Conflict 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.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2024.1469375/full#supplementary-material

References

Allen, G. C., Flores-Vergara, M. A., Krasynanski, S., Kumar, S., and Thompson, W. F. (2006). A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. *Nat. Protoc.* 1, 2320–2325. doi: 10.1038/nprot.2006.384

Bai, R., Zhang, Y., Jia, X., Fan, J., Hou, X., Wang, Y., et al. (2020). Isolation, characterization and immunomodulatory activity of oligosaccharides from Codonopsis pilosula. *J. Funct. Foods* 72, 104070. doi: 10.1016/j.jff.2020.104070

Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* 6, 11. doi: 10.1186/s13100-015-0041-9

Belton, J.-M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* 58, 268–276. doi: 10.1016/j.ymeth.2012.05.001

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi: 10.1093/nar/27.2.573

Birney, E., Clamp, M., and Durbin, R. (2004). Genewise and genomewise. *Genome Res.* 14, 988–995. doi: 10.1101/gr.1865504

Buchfink, B., Reuter, K., and Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* 18, 366–368. doi: 10.1038/s41592-021-01101-x

Cai, L. M., Luo, L., and Zeng, Y. H. (2014). Effects of polysaccharides from the root of Codonopsis pilosula (Dangshen) on physical fatigue induced by forced swimming. *Appl. Mechanics Mater.* 675, 1591–1594. doi: 10.4028/www.scientific.net/AMM.675-677

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinf.* 10, 421. doi: 10.1186/1471-2105-10-421

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973. doi: 10.1093/bioinformatics/btp348

Chan, P. P., Lin, B. Y., Mak, A. J., and Lowe, T. M. (2021). tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* 49, 9077–9096. doi: 10.1093/nar/gkab688

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890. doi: 10.1093/bioinformatics/bty560

Cheng, F., Wu, J., Cai, X., Liang, J., Freeling, M., and Wang, X. (2018). Gene retention, fractionation and subgenome differences in polyploid plants. *Nat. Plants* 4, 258–268. doi: 10.1038/s41477-018-0136-7

College, J. N. M. (1986). *Dictionary of traditional Chinese medicines* (Shanghai: Shanghai Science and Technology Publishing House).

Crowl, A. A., Miles, N. W., Visger, C. J., Hansen, K., Ayers, T., Haberle, R., et al. (2016). A global perspective on Campanulaceae: Biogeographic, genomic, and floral evolution. *Am. J. Bot.* 103, 233–245. doi: 10.3732/ajb.1500450

dos Reis, M. (2022). "Dating microbial evolution with MCMCtree," in *Environmental Microbial Evolution: Methods and Protocols*. Ed. H. Luo (Springer US, New York, NY), 3–22.

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., et al. (2017). *De novo* assembly of the Aedes aEgypti genome using Hi-C yields chromosome-length scaffolds. *Science* 356, 92–95. doi: 10.1126/science.aal3327

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S. P., Huntley, M. H., Lander, E. S., et al. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* 3, 95–98. doi: 10.1016/j.cels.2016.07.002

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinf.* 9, 18. doi: 10.1186/1471-2105-9-18

Emms, D. M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 238. doi: 10.1186/s13059-019-1832-y

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., et al. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci.* 117, 9451–9457. doi: 10.1073/pnas.1921046117

Gout, J.-F., Hao, Y., Johri, P., Arnaiz, O., Doak, T. G., Bhullar, S., et al. (2023). Dynamics of gene loss following ancient whole-genome duplication in the cryptic paramecium complex. *Mol. Biol. Evol.* 40, msad107. doi: 10.1093/molbev/msad107

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Guo, S., Zhang, J., Sun, H., Salse, J., Lucas, W. J., Zhang, H., et al. (2013). The draft genome of watermelon (Citrullus lanatus) and resequencing of 20 diverse accessions. *Nat. Genet.* 45, 51–58. doi: 10.1038/ng.2470

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using evidencemodeler and the program to assemble spliced alignments. *Genome Biol.* 9, R7. doi: 10.1186/gb-2008-9-1-r7

Hamilton, J. P., Godden, G. T., Lanier, E., Bhat, W. W., Kinser, T. J., Vaillancourt, B., et al. (2020). Generation of a chromosome-scale genome assembly of the insect-repellent terpenoid-producing Lamiaceae species, Callicarpa americana. *GigaScience* 9, giaa093. doi: 10.1093/gigascience/giaa093

He, L.-X., Zhang, Z.-F., Sun, B., Chen, Q.-H., Liu, R., Ren, J.-W., et al. (2016). Sea cucumber (Codonopsis pilosula) oligopeptides: immunomodulatory effects based on stimulating Th cells, cytokine secretion and antibody production. *Food Funct.* 7, 1208–1216. doi: 10.1039/C5FO01480H

Hu, J., Fan, J., Sun, Z., and Liu, S. (2020). NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* 36, 2253–2255. doi: 10.1093/bioinformatics/btz891

Hu, L., Xu, Z., Wang, M., Fan, R., Yuan, D., Wu, B., et al. (2019). The chromosome-scale reference genome of black pepper provides insight into piperine biosynthesis. *Nat. Commun.* 10, 4702. doi: 10.1038/s41467-019-12607-6

Jang, W., Kang, J.-N., Jo, I.-H., Lee, S.-M., Park, G.-H., and Kim, C.-K. (2023). The chromosome-level genome assembly of lance asiabell (Codonopsis lanceolata), a medicinal and vegetable plant of the Campanulaceae family. *Front. Genet.* 14. doi: 10.3389/fgene.2023.1100819

Jia, Y., Chen, S., Chen, W., Zhang, P., Su, Z., Zhang, L., et al. (2022). A chromosome-level reference genome of Chinese balloon flower (Platycodon grandiflorus). *Front. Genet.* 13. doi: 10.3389/fgene.2022.869784

Jiang, Y., Liu, Y., Guo, Q., Xu, C., Zhu, C., and Shi, J. (2016). Sesquiterpene glycosides from the roots of Codonopsis pilosula. *Acta Pharm. Sin. B* 6, 46–54. doi: 10.1016/j.apsb.2015.09.007

Kalvari, I., Nawrocki, E. P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M., et al. (2021). Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.* 49, D192–D200. doi: 10.1093/nar/gkaa1047

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth.4285

Kapusta, A., Suh, A., and Feschotte, C. (2017). Dynamics of genome size evolution in birds and mammals. *Proc. Natl. Acad. Sci.* 114, E1460–E1469. doi: 10.1073/pnas.1616702114

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* 37, 540–546. doi: 10.1038/s41587-019-0072-8

Kong, X., Zhang, Y., Wang, Z., Bao, S., Feng, Y., Wang, J., et al. (2023). Two-step model of paleohexaploidy, ancestral genome reshuffling and plasticity of heat shock response in Asteraceae. *Hortic. Res.* 10, uhad073. doi: 10.1093/hr/uhad073

Lee, D.-J., Choi, J.-W., Kang, J.-N., Lee, S.-M., Park, G.-H., and Kim, C.-K. (2023). Chromosome-scale genome assembly and triterpenoid saponin biosynthesis in Korean bellflower (Platycodon grandiflorum). *Int. J. Mol. Sci.* 24, 6534. doi: 10.3390/ijms24076534

Leebens-Mack, J. H., Barker, M. S., Carpenter, E. J., Deyholos, M. K., Gitzendanner, M. A., Graham, S. W., et al. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574, 679–685. doi: 10.1038/s41586-019-1693-2

Letunic, I., and Bork, P. (2024). Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res.* 52, W78–W82. doi: 10.1093/nar/gkae268

Lin, L.-C., Tsai, T.-H., and Kuo, C.-L. (2013). Chemical constituents comparison of Codonopsis tangshenCodonopsis pilosula var. modesta and Codonopsis pilosula. *Natural Product Res.* 27, 1812–1815. doi: 10.1080/14786419.2013.778849

Liu, X., Gong, Q., Zhao, C., Wang, D., Ye, X., Zheng, G., et al. (2023). Genome-wide analysis of cytochrome P450 genes in Citrus clementina and characterization of a CYP gene encoding flavonoid 3′-hydroxylase. *Hortic. Res.* 10, uhac283. doi: 10.1093/hr/uhac283

Liu, Y., Wang, S., Li, L., Yang, T., Dong, S., Wei, T., et al. (2022). The Cycas genome and the early evolution of seed plants. *Nat. Plants* 8, 389–401. doi: 10.1038/s41477-022-01129-7

Ma, Y., Liu, D., Wariss, H. M., Zhang, R., Tao, L., Milne, R. I., et al. (2022). Demographic history and identification of threats revealed by population genomic analysis provide insights into conservation for an endangered maple. *Mol. Ecol.* 31, 767–779. doi: 10.1111/mec.16289

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., and Ravikesavan, R. (2013). Gene duplication as a major force in evolution. *J. Genet.* 92, 155–161. doi: 10.1007/s12041-013-0212-8

Mandáková, T., and Lysak, M. A. (2018). Post-polyploid diploidization and diversification through dysploid changes. *Curr. Opin. Plant Biol.* 42, 55–65. doi: 10.1016/j.pbi.2018.03.001

Manni, M., Berkeley, M. R., Seppey, M., and Zdobnov, E. M. (2021). BUSCO: assessing genomic data quality and beyond. *Curr. Protoc.* 1, e323. doi: 10.1002/cpz1.323

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770. doi: 10.1093/bioinformatics/btr011

Mendes, F. K., Vanderpool, D., Fulton, B., and Hahn, M. W. (2021). CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* 36, 5516–5518. doi: 10.1093/bioinformatics/btaa1022

Meng, Y., Xu, Y., Chang, C., Qiu, Z., Hu, J., Wu, Y., et al. (2020). Extraction, characterization and anti-inflammatory activities of an inulin-type fructan from Codonopsis pilosula. *Int. J. Biol. Macromol.* 163, 1677–1686. doi: 10.1016/j.ijbiomac.2020.09.117

Nawrocki, E. P., and Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29, 2933–2935. doi: 10.1093/bioinformatics/btt509

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3, 418–426. doi: 10.1093/oxfordjournals.molbev.a040410

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Ou, S., Chen, J., and Jiang, N. (2018). Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* 46, e126. doi: 10.1093/nar/gky730

Ou, S., and Jiang, N. (2018). LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* 176, 1410–1422. doi: 10.1104/pp.17.01310

Ou, S., and Jiang, N. (2019). LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mobile DNA* 10, 48. doi: 10.1186/s13100-019-0193-0

Pellicer, J., Hidalgo, O., Dodsworth, S., and Leitch, I. J. (2018). Genome size diversity and its impact on the evolution of land plants. *Genes* 9, 88. doi: 10.3390/genes9020088

Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., and Finn, R. D. (2018). HMMER web server: 2018 update. *Nucleic Acids Res.* 46, W200–W204. doi: 10.1093/nar/gky448

Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., et al. (2019). Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biol.* 20, 38. doi: 10.1186/s13059-019-1650-2

Ranallo-Benavidez, T. R., Jaron, K. S., and Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* 11, 1432. doi: 10.1038/s41467-020-14998-3

Ren, R., Wang, H. F., Guo, C. C., Zhang, N., Zeng, L. P., Chen, Y. M., et al. (2018). Widespread whole genome duplications contribute to genome complexity and species diversity in angiosperms. *Mol. Plant* 11, 414–428. doi: 10.1016/j.molp.2018.01.002

Rhie, A., Walenz, B. P., Koren, S., and Phillippy, A. M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* 21, 245. doi: 10.1186/s13059-020-02134-9

Roach, M. J., Schmidt, S. A., and Borneman, A. R. (2018). Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinf.* 19, 460. doi: 10.1186/s12859-018-2485-7

Robinson, J. T., Turner, D., Durand, N. C., Thorvaldsdóttir, H., Mesirov, J. P., and Aiden, E. L. (2018). Juicebox.js provides a cloud-based visualization system for Hi-C data. *Cell Syst.* 6, 256–258.e251. doi: 10.1016/j.cels.2018.01.001

Sato, S., Tabata, S., Hirakawa, H., Asamizu, E., Shirasawa, K., Isobe, S., et al. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485, 635–641. doi: 10.1038/nature11119

Shi, X., Cao, S., Wang, X., Huang, S., Wang, Y., Liu, Z., et al. (2023). The complete reference genome for grapevine (*Vitis vinifera* L.) genetics and breeding. *Hortic. Res.* 10, uhad061. doi: 10.1093/hr/uhad061

Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics* 24, 637–644. doi: 10.1093/bioinformatics/btn013

Sun, P., Jiao, B., Yang, Y., Shan, L., Li, T., Li, X., et al. (2022). WGDI: a user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. *Mol. Plant* 15, 1841–1851. doi: 10.1016/j.molp.2022.10.018

Sun, P., Lu, Z., Wang, Z., Wang, S., Zhao, K., Mei, D., et al. (2024). Subgenome-aware analyses reveal the genomic consequences of ancient allopolyploid hybridizations throughout the cotton family. *Proc. Natl. Acad. Sci.* 121, e2313921121. doi: 10.1073/pnas.2313921121

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34, W609–W612. doi: 10.1093/nar/gkl315

Tang, H., Krishnakumar, V., Zeng, X., Xu, Z., Taranto, A., Lomas, J. S., et al. (2024). JCVI: A versatile toolkit for comparative genomics analysis. *iMeta* 3, e211. doi: 10.1002/imt2.211

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinf.* 25, 4.10.11–14.10.14. doi: 10.1002/0471250953.bi0410s25

Tu, L., Su, P., Zhang, Z., Gao, L., Wang, J., Hu, T., et al. (2020). Genome of Tripterygium wilfordii and identification of cytochrome P450 involved in triptolide biosynthesis. *Nat. Commun.* 11, 971. doi: 10.1038/s41467-020-14776-1

Vo, T. P., Ho, M. T., Nguyen Nguyen, P. U., Pham, N. D., Truong, K. V., Yen Nguyen, T. H., et al. (2024). Extracting phenolics, flavonoids, and terpenoids from Codonopsis pilosula using green solvents. *Sustain. Chem. Pharm.* 37, 101395. doi: 10.1016/j.scp.2023.101395

Wan, T., Liu, Z., Leitch, I. J., Xin, H., Maggs-Kölling, G., Gong, Y., et al. (2021). The Welwitschia genome reveals a unique biology underpinning extreme longevity in deserts. *Nat. Commun.* 12, 4247. doi: 10.1038/s41467-021-24528-4

Wang, D., Zheng, Z., Li, Y., Hu, H., Wang, Z., Du, X., et al. (2021). Which factors contribute most to genome size variation within angiosperms? *Ecol. Evol.* 11, 2660–2668. doi: 10.1002/ece3.7222

Wendel, J. F., Jackson, S. A., Meyers, B. C., and Wing, R. A. (2016). Evolution of plant genome architecture. *Genome Biol.* 17, 37. doi: 10.1186/s13059-016-0908-1

Willing, E.-M., Rawat, V., Mandáková, T., Maumus, F., James, G. V., Nordström, K. J. V., et al. (2015). Genome expansion of *Arabis alpina* linked with retrotransposition and reduced symmetric DNA methylation. *Nat. Plants* 1, 14023. doi: 10.1038/nplants.2014.23

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation* 2, 100141. doi: 10.1016/j.xinn.2021.100141

Wu, B., Yu, Q., Deng, Z., Duan, Y., Luo, F., and Gmitter, F. Jr. (2023). A chromosome-level phased genome enabling allele-level studies in sweet orange: a case study on citrus Huanglongbing tolerance. *Hortic. Res.* 10, uhac247. doi: 10.1093/hr/uhac247

Xu, C., Liu, Y., Yuan, G., and Guan, M. (2012). The contribution of side chains to antitumor activity of a polysaccharide from Codonopsis pilosula. *Int. J. Biol. Macromol.* 50, 891–894. doi: 10.1016/j.ijbiomac.2012.01.013

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Ye, T., Li, S., Li, Y., Xiao, S., and Yuan, D. (2024). Impact of polyploidization on genome evolution and phenotypic diversity in oil-tea Camellia. *Ind. Crops Products* 218, 118928. doi: 10.1016/j.indcrop.2024.118928

Zhang, T., Qiao, Q., Du, X., Zhang, X., Hou, Y., Wei, X., et al. (2022). Cultivated hawthorn (Crataegus pinnatifida var. major) genome sheds light on the evolution of Maleae (apple tribe). *J. Integr. Plant Biol.* 64, 1487–1501. doi: 10.1111/jipb.13318

Zhu, Q., Wang, Y., Yao, N., Ni, X., Wang, C., Wang, M., et al. (2023). Chromosome-level genome assembly of an endangered plant Prunus mongolica using PacBio and Hi-C technologies. *DNA Res.* 30, dsad012. doi: 10.1093/dnares/dsad012

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., and Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics* 29, 2669–2677. doi: 10.1093/bioinformatics/btt476

Zou, Y.-F., Zhang, Y.-Y., Paulsen, B. S., Fu, Y.-P., Huang, C., Feng, B., et al. (2020). Prospects of Codonopsis pilosula polysaccharides: Structural features and bioactivities diversity. *Trends Food Sci. Technol.* 103, 1–11. doi: 10.1016/j.tifs.2020.06.012

Zuntini, A. R., Carruthers, T., Maurin, O., Bailey, P. C., Leempoel, K., Brewer, G. E., et al. (2024). Phylogenomics and the rise of the angiosperms. *Nature* 629, 843–850. doi: 10.1038/s41586-024-07324-0

Zwaenepoel, A., Li, Z., Lohaus, R., and Van de Peer, Y. (2018). Finding evidence for whole genome duplications: A reappraisal. *Mol. Plant* 12, 133–136. doi: 10.1016/j.molp.2018.12.019