Chromosome-level assembly of the *Clinopodium gracile* genome

Yubang Gao*

School of Life Sciences, Nanyang Normal University, Nanyang, Henan, China
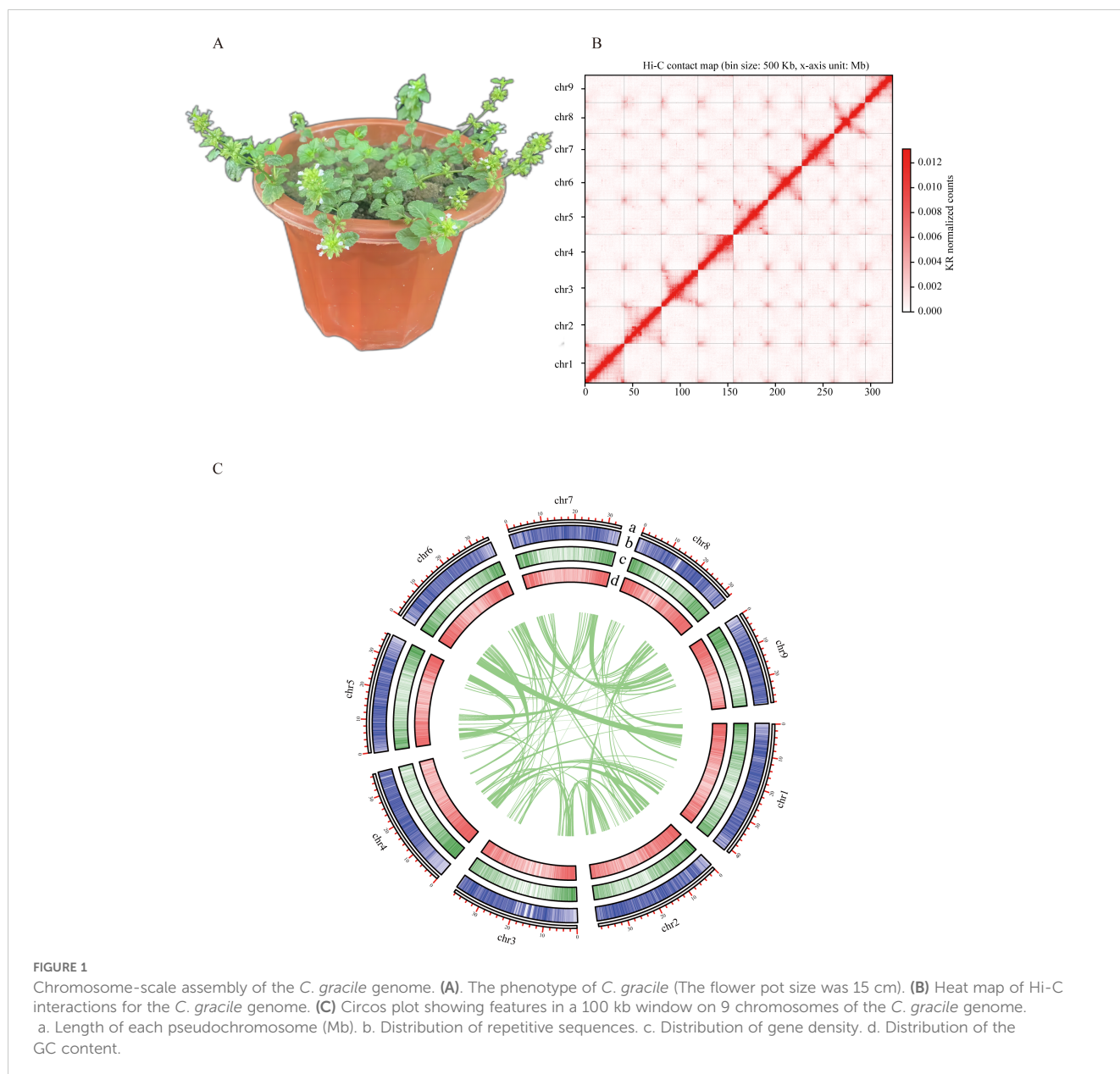
Clinopodium gracile is an important medicinal herb in the Lamiaceae family. This species lacks corresponding genomic resources, which significantly limits the study of its active compound synthesis pathways, breeding practices, and assessment of natural genetic variations. We assembled the chromosomal-level genome of *C. gracile* using Oxford Nanopore (ONT) technology and Hi-C sequence. The assembled genome is 307.3 Mb in size and consists of 9 chromosomes. The scaffold N50 was 36.3 Mb. The BUSCO completeness (Embryophyta_db10) of the genome was 97.2%. The genome annotates 40,083 protein coding genes. *C. gracile* and *S. miltiorrhiza* diverged approximately 30.615 million years ago. *C. gracile* has not undergone recent species-specific WGD events. A high proportion of young LTRs indicates a recent transposable element (TE) transposition burst in *C. gracile*.

KEYWORDS

genome, chinese herbal medicine, Clinopodium gracile, nanopore sequence, Hi-C assembly

1 Introduction

C. gracile (Figure 1A) is a medicinal plant belonging to the genus Clinopodium in the Lamiaceae family (Dai et al., 1984). The Clinopodium genus comprises 20 species, most of which are medicinal plants (Yao et al., 2020). The triterpenoid saponins of *C. gracile* exhibit various pharmacological effects, such as anti-inflammatory (Park et al., 2010), cardioprotection (Hu et al., 2017), and anti-tumor characteristics (Dzhambazov et al., 2002). Additionally, it exhibits insecticidal activities (Chen et al., 2013). Research on *C. gracile* involves transcriptomics of different tissues analysis (Zhao et al., 2020; Shan et al., 2020). Moreover, studies on species like *Clinopodium chinese* include transcriptomics analysis (Shi et al., 2019) and microRNA analysis (Xu et al., 2022). *C. gracile* and other plants in the same genus lack corresponding

FIGURE 1
Chromosome-scale assembly of the *C. gracile* genome. **(A)**. The phenotype of *C. gracile* (The flower pot size was 15 cm). **(B)** Heat map of Hi-C interactions for the *C. gracile* genome. **(C)** Circos plot showing features in a 100 kb window on 9 chromosomes of the *C. gracile* genome. a. Length of each pseudochromosome (Mb). b. Distribution of repetitive sequences. c. Distribution of gene density. d. Distribution of the GC content.

reference genomes, significantly hindering the study of their active compound synthesis pathways, breeding practices, and assessment of natural genetic variations.

With continuous improvements in DNA sequencing technologies, assembling chromosome-level genomes is becoming increasingly feasible (Kong et al., 2023). Nanopore sequencing, known for its long-read capability, offers distinct advantages in genome assembly (MacKenzie and Argyropoulos, 2023). Here, we employed ONT nanopore sequencing and Hi-C sequence to assemble the *C. gracile* genome. The assembly's contig N50 values were 36.3 Mb. Comparative genomic analysis indicated that the *C. gracile* genome had undergone a TE insertion burst. The assembled genome with gene annotations is the first reference genome for this species and the genus. The assembled genome in this study will facilitate research on the *C. gracile* genome, metabolic engineering, and the improvement of elite cultivars.

2 Materials and methods

2.1 Material collection and genome sequencing

The plants used for genome sequencing were cultivated under laboratory conditions: 25℃, 3000 lx, and a photoperiod of 16 hours light to 8 hours dark. High molecular weight DNA (HMW DNA) was extracted for subsequent library construction. Genomic DNA was extracted using the Qiagen MagAttract HMW DNA Mini Kit, following the manufacturer's protocol. The Hi-C libraries were prepared by chromatin crosslinking, restricted enzyme (MboI) digestion, end filling and biotinylation tagging, DNA purification and shearing. All of the prepared DNA fragments were processed into paired-end sequencing libraries. Sequencing was performed on the DNBSEQ-T7 and PromethION platforms.

2.2 Genome survey

We used fastp (Chen et al., 2018) version 0.20.1 to trim the raw reads. Using the trimmed data, we employed Jellyfish (Marçais and Kingsford, 2011) version 2.3.0 to calculate the K-mer distribution histogram. Genome size, heterozygosity, and repeat rate were estimated using GenomeScope 2.0 (Ranallo-Benavidez et al., 2020). Genomic ploidy was analyzed using Smudgeplot (Ranallo-Benavidez et al., 2020).

2.3 Genome assembly and gene annotation

The genome was assembled using NextDenovo (Hu et al., 2023) version 2.5.2. The assembled sequences were polished four times using NextPolish (Hu et al., 2020) version 1.4.1 with short read. Repetitive elements in the genome were annotated using EDTA version 2.0.1 (Ou et al., 2019). Gene prediction was conducted using Funannotate version 1.8.16 (https://github.com/nextgenusfs/funannotate), integrating *de novo* prediction, homology prediction, and transcriptome sequencing data. Functional annotation was performed using DIAMOND (Buchfink et al., 2015) version 2.0.14.152 for protein BLAST against EggNOG/SwissProt/NR/TAIR databases. tRNAs, rRNAs, miRNAs, and snRNAs were identified using Infernal (Nawrocki and Eddy, 2013) version 1.1. The completeness of the genome assembly and protein-coding genes were evaluated using BUSCO (Simão et al., 2015) version 5.2.2. The assembly quality was assessed by mapping short-read data to the assembled genome using Bowtie2 (Langmead and Salzberg, 2012) version 2.4.4. Long-read data was mapped to the assembled genome using Minimap2 (Li, 2018) version 2.24-r1122.

2.4 Phylogenetic analysis

Protein sequences from *C. gracile* and ten other flowering plants (*Oryza sativa*, *Amborella trichopoda*, *Arabidopsis thaliana*, *Coffea canephora*, *Theobroma cacao*, *Vitis vinifera*, *Salvia miltiorrhiza*, *Leonurus japonicus*, *Tectona grandis*, and S*olanum lycopersicum*) were utilized to create a phylogenetic tree. OrthoVenn3 (Sun et al., 2023) supported analyses of phylogenetic and gene family expansions and contractions. The process entailed using OrthoMCL (Li et al., 2003) for identifying homologous proteins and unique genes. FastTree2 (Price et al., 2010) version 2.1.7 was employed to develop the phylogenetic tree using the JTT+CAT model. SH tests verified node accuracy. Divergence among species was calculated using the r8s tool (Sanderson, 2003) version 1.81 with known divergence times between *A. thaliana* and *C. canephora*, *A. thaliana* and *V. vinifera*, *S. lycopersicum* and *T. cacao*, *S. miltiorrhiza* and *L. japonicus*. Additionally, gene family expansions and contractions were evaluated with CAFE5 (Mendes et al., 2020), applying a stochastic birth-and-death model. Assess the statistical significance via conditional likelihood, with P-values <= 0.01 indicating significant findings.

2.5 Duplicated gene analysis

MCScanX (Wang et al., 2012) was used to detect synteny and collinearity within and between species. Duplicated genes originating from WGDs were extracted from collinear regions. The downstream analysis script 'duplicate_gene_classifier' from MCScanX was utilized to categorize types of duplicated genes. Based on codon alignments using the YN substitution model, the four-fold degenerate transversion (4DTv) distances were calculated between orthologous and paralogous gene pairs within and between species. GO enrichment analysis was performed using ClusterProfiler (Wu et al., 2021) version 4.0. Circos are plotted using the TBtools (Chen et al., 2020) circos function. The link size parameter under the link region config setting was set to 0.1.

2.6 RNA-Seq analysis

Previously published RNA-Seq data for *C. gracile* roots, stems, leaves, and flowers (Shan et al., 2020) were downloaded. The downloaded RNA-seq reads were mapped to the *C. gracile* genome using HISAT2 (Kim et al., 2019) version 2.2.1. Gene expression levels were quantified by calculating FPKM values using StringTie2 (Kovaka et al., 2019) version 2.2.1.

3 Data

3.1 Genome assembly

32.26 Gb short-read data, 32.46 Gb Nanopore long-read data and 43.2 Gb Hi-C data were generated (Supplementary Table S1). Genome survey using short-read data revealed a genome size of 269.73 MB, with repeat elements constituting 36.9% and heterozygosity of 0.27% (Supplementary Figure S1). Genomic ploidy analysis predicts the *C. gracile* genome to be diploid (Supplementary Figure S2). The long-read assembly resulted in a 307 Mb genome comprising 31 contigs, with an N50 of 28.2 Mb. Post-scaffolding with Hi-C data yielded 9 pseudochromosomes with an N50 of 36.3 Mb (Figure 1B). The pseudochromosome ranges from 39.2 Mb to 26.9 Mb, covering 99.7% of the genome. Chromosomes were numbered in descending order of size. For the genome assembly, the BUSCO completeness was 97.21% (Supplementary Table S2). The mapping rates of short-read and long-read genomic data to the unmasked genome were 88.72% and 99.38%, respectively. The mapping rate of previously published RNA-Seq data (Shan et al., 2020) was 93.90%. These results indicated good assembly quality.

3.2 Gene prediction and gene annotation

The genome contains 51.39% repetitive sequences, with Type I transposable elements (TEs) comprising 36.35% and Type II TEs 15.04%. Within Type I TEs (LTR-RTs), the highest proportion is
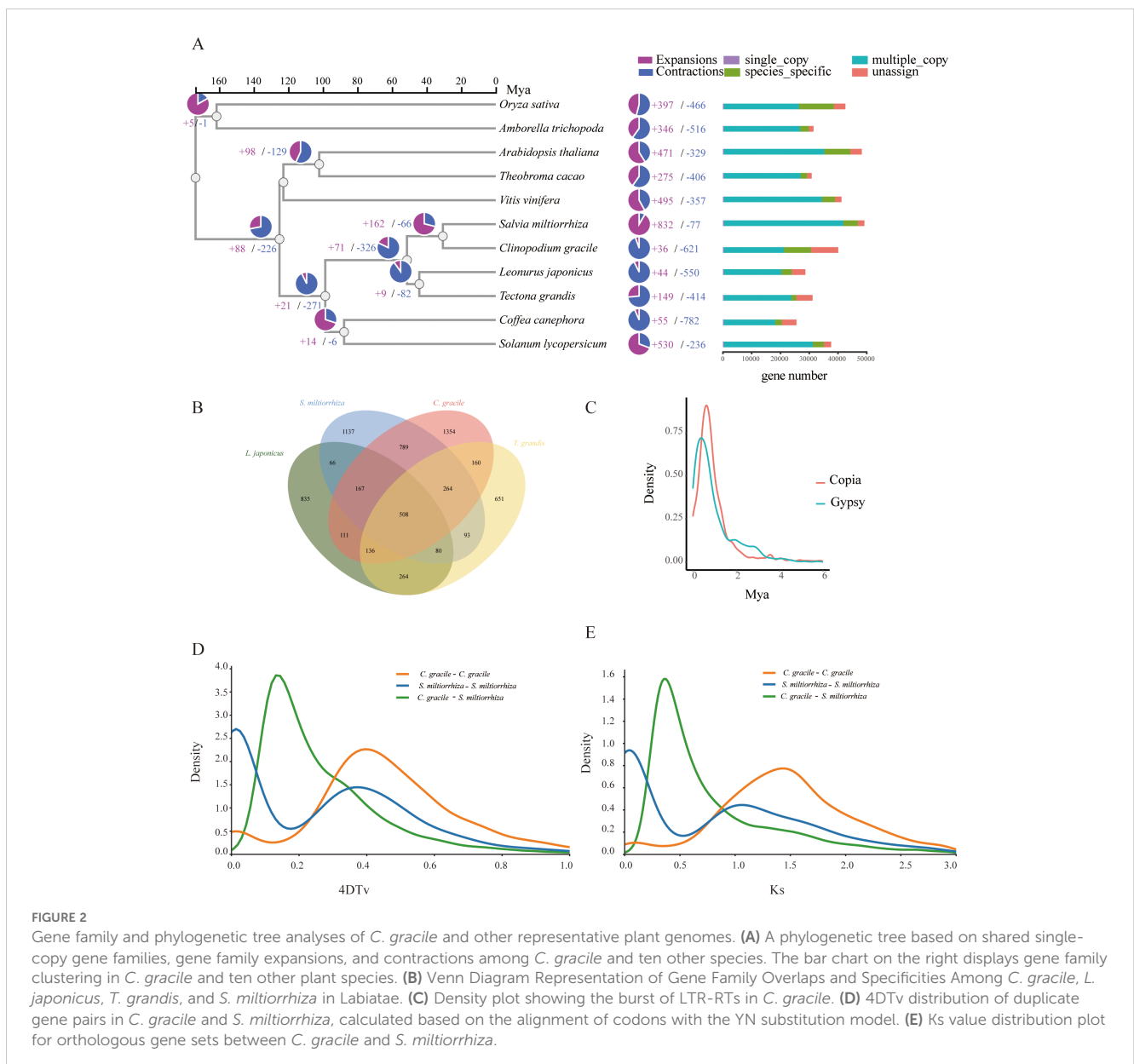
Gypsy (21.92%), followed by Copia (12.47%). Consistent with most plants, LTR-RT represents the most prevalent elements in the *C. gracile* genome. In *C. gracile*, the Gypsy (21.92%) and Copia (12.47%) retrotransposon families exhibit a slight contraction compared to *Salvia miltiorrhiza* (Gypsy: 29.83%, Copia: 14.77%) (Song et al., 2020).

After masking repetitive sequences, we predicted 40,083 protein-coding genes (Figure 1C). The average coding sequence (CDS) length of genes is 978 bp (Supplementary Table S3). Genes contain 4.33 exons on average. 34,193 (85.3%) predicted genes could be annotated in public databases (EggNOG, NR, Swiss-Prot, and TAIR). These results indicate that the *C. gracile* genome assembly is high quality and nearly complete. Terpene synthesis-related genes in *T. grandis* (Zhao et al., 2019) were used for homology prediction. A total of 42 genes of terpene synthesis-related pathways were predicted. Of these, 14 were involved in the mevalonate (MVA) pathway and 28 in the methylerythritol phosphate (MEP) pathway (Supplementary Table

S6). Additionally, we predicted 732 rRNAs, 634 tRNAs, 492 miRNAs, and 1009 snRNAs.

3.3 Comparative genomic analysis of *C. gracile* with other plants

Combining the protein sequences of *C. gracile* and ten other angiosperms yielded 40,083 proteins (Supplementary Table S4). Clustering identified 34,011 gene families, including 319 single-copy gene families (Supplementary Table S5). These single-copy genes were used to construct a phylogenetic tree, incorporating known divergence times. *C. gracile* and *S. miltiorrhiza* diverged approximately 30.615 million years ago (MYA) (Figure 2A). Synteny analysis between *C. gracile* and *S. miltiorrhiza* revealed limited collinearity (Supplementary Figure S3), indicating significant genomic changes since their divergence. In *C. gracile*, there are expansions in 36 gene families

FIGURE 2
Gene family and phylogenetic tree analyses of *C. gracile* and other representative plant genomes. **(A)** A phylogenetic tree based on shared single-copy gene families, gene family expansions, and contractions among *C. gracile* and ten other species. The bar chart on the right displays gene family clustering in *C. gracile* and ten other plant species. **(B)** Venn Diagram Representation of Gene Family Overlaps and Specificities Among *C. gracile*, *L. japonicus*, *T. grandis*, and *S. miltiorrhiza* in Labiatae. **(C)** Density plot showing the burst of LTR-RTs in *C. gracile*. **(D)** 4DTv distribution of duplicate gene pairs in *C. gracile* and *S. miltiorrhiza*, calculated based on the alignment of codons with the YN substitution model. **(E)** Ks value distribution plot for orthologous gene sets between *C. gracile* and *S. miltiorrhiza*.

and contractions in 621 gene families. The expanded genes are primarily enriched in categories related to environmental stress, such as response to growth hormone, toxic substances, and cadmium ion (Supplementary Figure S4). A comparison across four Lamiaceae species showed 508 conserved gene families, suggesting high conservation within the family (Figure 2B). *C. gracile* and *S. miltiorrhiza* have retained the most species-specific gene families, which may be functionally unique.

TEs play a significant role in genome evolution. In the *C. gracile* genome, the LTR-RT families were analyzed (Figure 2C). There was one peak (Gypsy or Copia) of LTR-RT amplification within the last 1 million years, indicating a recent burst in LTR-RT amplification in its genome. We identified 1,119 LTR-RTs (87.4%) with insertion times less than 2 MYA. The high proportion of young LTR-RTs suggests that TE transposition has been actively shaping the recent evolutionary history of *C. gracile*.

To further investigate the evolutionary differences between *C. gracile* and *S. miltiorrhiza*, we analyzed the four-fold degenerate transversion (4DTv) rates among orthologous gene pairs within and between species. The peak 4DTv distance of 0.17 corresponds to the speciation event that separated *C. gracile* and *S. miltiorrhiza* (Figure 2D). The Ks distribution plot for syntenic genes between *C. gracile* and *S. miltiorrhiza* shows similar trends (Figure 2E). The duplicated genes in the genome were categorized, resulting in 6,481 whole-genome duplications (WGDs), 3,738 tandem duplications, 2,529 proximal, 18,144 dispersed, and 9,722 singleton duplications. Tandem duplications are particularly enriched in the secondary metabolite biosynthetic process, response to toxic substances, and toxin metabolic process (Supplementary Figure S5), suggesting their role in metabolizing secondary metabolites and toxic substances. We find 225 positively selected genes (Ka/Ks > 1) and 2,022 negatively selected genes (Ka/Ks < 1) (Supplementary Figure S6). Positively selected genes were enriched in cellular respiration, heat acclimation, and positive regulation of auxin-mediated signaling pathways, indicating selection by harsh environmental conditions (Supplementary Figure S7).

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/, SRR28508814, SRR28508813, SRR29849768 https://figshare.com/, https://figshare.com/s/579e58be9da2ccbcb192.

Conflict of interest

The author declares 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

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2024.1489102/full#supplementary-material

References

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp.2020.06.009

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

Chen, X. B., Liu, X. C., Zhou, L., and Liu, Z. L. (2013). Essential oil composition and larvicidal activity of Clinopodium gracile (Benth) Matsum (Labiatae) aerial parts against the Aedes albopictus mosquito. *Trop. J. Pharm. Res.* 12, 799–804. doi: 10.4314/tjpr.v12i5.21

Dai, J., Shi, D., and Zhang, H. (1984). Morphology, anatomy and chemical constituents of five species of Clinopodium. *Acta Pharm. Sin. B* 19 (6), 425–430.

Dzhambazov, B., Daskalova, S., Monteva, A., and Popov, N. (2002). *In vitro* screening for antitumour activity of Clinopodium vulgare L.(Lamiaceae) extracts. *Biol. Pharm. Bull.* 25, 499–504. doi: 10.1248/bpb.25.499

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, J., Wang, Z., Sun, Z., Hu, B., Ayoola, A. O., Liang, F., et al. (2023). An efficient error correction and accurate assembly tool for noisy long reads. *Genome Biol.* 25 (1), 107. doi: 10.1186/s13059-024-03252-4

Hu, Y.-X., Zhang, W., Zhang, W., Zhu, Y.-D., Ma, G.-X., Zhu, N.-L., et al. (2017). Oleanane triterpene saponins with cardioprotective activity from Clinopodium polycephalum. *J. Asian Natural Products Res.* 19, 697–703. doi: 10.1080/10286020.2016.1254199

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37, 907–915. doi: 10.1038/s41587-019-0201-4

Kong, W., Wang, Y., Zhang, S., Yu, J., and Zhang, X. (2023). Recent advances in assembly of complex plant genomes. *Genomics Proteomics Bioinf.* 21, 427–439. doi: 10.1016/j.gpb.2023.04.004

Kovaka, S., Zimin, A. V., Pertea, G. M., Razaghi, R., Salzberg, S. L., and Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* 20, 1–13. doi: 10.1186/s13059-019-1910-1

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189. doi: 10.1101/gr.1224503

MacKenzie, M., and Argyropoulos, C. (2023). An introduction to nanopore sequencing: past, present, and future considerations. *Micromachines* 14, 459. doi: 10.3390/mi14020459

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. (2020). CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* 36, 5516–5518. doi: 10.1093/bioinformatics/btaa1022

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

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R., Hellinga, A. J., et al. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* 20, 1–18. doi: 10.1186/s13059-019-1905-y

Park, S.-B., Kim, S.-H., Suk, K., Lee, H.-S., Kwon, T. K., Ju, M.-G., et al. (2010). Clinopodium gracile inhibits mast cell-mediated allergic inflammation: involvement of calcium and nuclear factor-κ B. *Exp. Biol. Med.* 235, 606–613. doi: 10.1258/ebm.2010.009292

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS One* 5, e9490. doi: 10.1371/journal.pone.0009490

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

Sanderson, M. J. (2003). r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19, 301–302. doi: 10.1093/bioinformatics/19.2.301

Shan, C., Wang, C., Zhang, S., Shi, Y., Ma, K., Yang, Q., et al. (2020). Transcriptome analysis of Clinopodium gracile (Benth.) Matsum and identification of genes related to Triterpenoid Saponin biosynthesis. *BMC Genomics* 21, 1–16. doi: 10.1186/s12864-020-6454-y

Shi, Y., Zhang, S., Peng, D., Wang, C., Zhao, D., Ma, K., et al. (2019). Transcriptome analysis of Clinopodium chinense (Benth.) O. Kuntze and identification of genes involved in Triterpenoid Saponin biosynthesis. *Int. J. Mol. Sci.* 20, 2643. doi: 10.3390/ijms20112643

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Song, Z., Lin, C., Xing, P., Fen, Y., Jin, H., Zhou, C., et al. (2020). A high-quality reference genome sequence of Salvia miltiorrhiza provides insights into tanshinone synthesis in its red rhizomes. *Plant Genome* 13, e20041. doi: 10.1002/tpg2.20041

Sun, J., Lu, F., Luo, Y., Bie, L., Xu, L., and Wang, Y. (2023). OrthoVenn3: an integrated platform for exploring and visualizing orthologous data across genomes. *Nucleic Acids Res.* 51, W397–W403. doi: 10.1093/nar/gkad313

Wang, Y., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40, e49–e49. doi: 10.1093/nar/gkr1293

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 (3). doi: 10.1016/j.xinn.2021.100141

Xu, J.-Y., Shan, C.-M., Shan, T.-Y., Zhao, L.-Q., Ma, K.-L., and Wu, J.-W. (2022). Analysis of microRNAs and their target genes in Clinopodium chinense (Benth.) O. Kuntze using small RNA sequencing. *Plant Sci. J.* 40, 216–228. doi: 10.11913/PSJ.2095-0837.2022.20216

Yao, L., Li-Min, L., Yuan-Gen, X., Xue-Bin, S., Huan, L., and Si-Hui, N. (2020). Research progress on genus Clinopodium. *China J. Chin. Materia Med.* 45, 4349–4357. doi: 10.19540/j.cnki.cjcmm.20200604.601

Zhao, D., Hamilton, J. P., Bhat, W. W., Johnson, S. R., Godden, G. T., Kinser, T. J., et al. (2019). A chromosomal-scale genome assembly of Tectona grandis reveals the importance of tandem gene duplication and enables discovery of genes in natural product biosynthetic pathways. *Gigascience* 8. doi: 10.1093/gigascience/giz005

Zhao, L.-Q., Shan, C.-M., Zhang, S.-X., Shi, Y.-Y., Ma, K.-L., and Wu, J.-W. (2020). Identification of key enzyme genes involved in anthocyanin synthesis pathway in Clinopodium gracile by transcriptome analysis. *Bull. Bot. Res.* 40 (6), 886–896. doi: 10.7525/j.issn.1673-5102.2020.06.011