



OPEN

DATA DESCRIPTOR

Chromosome-level assembly and evolution analysis of the *Trichosanthes truncata* genome

Ying Hu^{1,2}, Xiaomei Wei^{1,2}, Zhuannan Chu³, Fan Wei^{1,2}, Yude Peng¹, Baoyou Huang¹, Ling Dong³, Kunhua Wei^{1,2}✉ & Weiwen Li³✉

Trichosanthes truncata C. B. Clarke, an important medicinal plant, is a dioecious plant belonging to the Cucurbitaceae family. This study presents a chromosomal-level reference genome assembly for *T. truncata*. Through the integration of PacBio high-fidelity sequencing and high-throughput chromosome conformation capture technology, a final genome sequence of 637.41 Mb was assembled, with an N50 of 57.24 Mb and consisting of 11 pseudochromosomes. Additionally, 97.21 Mb of repetitive sequences and 36,172 protein-coding genes were annotated. This high-quality genome assembly is of utmost significance for studying the molecular mechanisms underlying the biosynthesis of bioactive compounds. Furthermore, this study provided valuable insights into plant comparative genomics research.

Background & Summary

Trichosanthes, which belongs to the family Cucurbitaceae, is one of the largest genera, most of which are annual or perennial herbs distributed in tropical Asia and Australia¹. It has been reported that *Trichosanthes* ranks among the top five genera with considerable potential for discovery of anticancer agents out of the 145 families of plants². *Trichosanthes* has been found to possess significant medicinal properties, including anti-inflammatory, antioxidant, hypoglycemic, and anticancer effects^{1,2}.

Trichosanthes truncata C. B. Clarke, commonly known as the pointed leaf gourd, is a dioecious perennial vine herb belonging to the Cucurbitaceae family³ (Fig. 1a). *T. truncata* contains abundant organic Se and beneficial bioactive substances such as trichosantin and exhibits diverse effects such as coronary artery dilation, increased blood flow, reduced heart rate, enhanced hypoxia tolerance, acute myocardial ischemia resistance, and platelet aggregation inhibition^{3,4}. Furthermore, its toxicity is lower than that of other *Trichosanthes* species, emphasizing its significant medicinal value^{3,4}. In addition, most *Trichosanthes* species, including *T. truncata*, are dioecious⁵. In contrast, another *Trichosanthes* species, the snake gourd (*Trichosanthes anguina* L.), which belongs to the same genus and a fully assembled genome sequence is available, is monoecious and diclinous^{6,7}.

Given the limited genomic resources of *T. truncata*, comprehensive insights into the genetic characteristics and functions associated with the production of bioactive compounds remain limited. We employed PacBio high-fidelity (HiFi) sequencing and high-throughput chromosome conformation capture (Hi-C) techniques to obtain chromosome-level reference genome sequences for *T. truncata*. In this study, we generated a high-quality genome assembly of *T. truncata*, which will greatly facilitate the elucidation of the molecular mechanisms underlying the biosynthesis of beneficial compounds with medicinal value in *T. truncata*, the exploration of comparative genomic research in the Cucurbitaceae family, thus providing a vital genomic resource and a foundation for further investigations.

Methods

Sample preparation and DNA sequencing. The voucher specimen of *T. truncata* was stored at Guangxi Botanical Garden of Medicinal Plants (<http://www.gxyyzwy.com/>), Ying Hu, hying@gxyyzwy.com), with a voucher number of YY10816. *T. truncata* fresh leaves were collected from the introduction and preservation nurseries in the Guangxi Botanical Garden of Medical Plants, and the samples were stored at 4 °C before being

¹National Center for Traditional Chinese Medicine Inheritance and Innovation, Guangxi Botanical Garden of Medicinal Plants, Nanning, 530023, China. ²Guangxi Key Laboratory of Medicinal Resources Protection and Genetic Improvement, Guangxi Botanical Garden of Medicinal Plants, Nanning, 530023, China. ³Institute of Horticulture, Anhui Academy of Agricultural Sciences, Hefei, 230001, China. ✉e-mail: divinekh@163.com; doulala8404@126.com

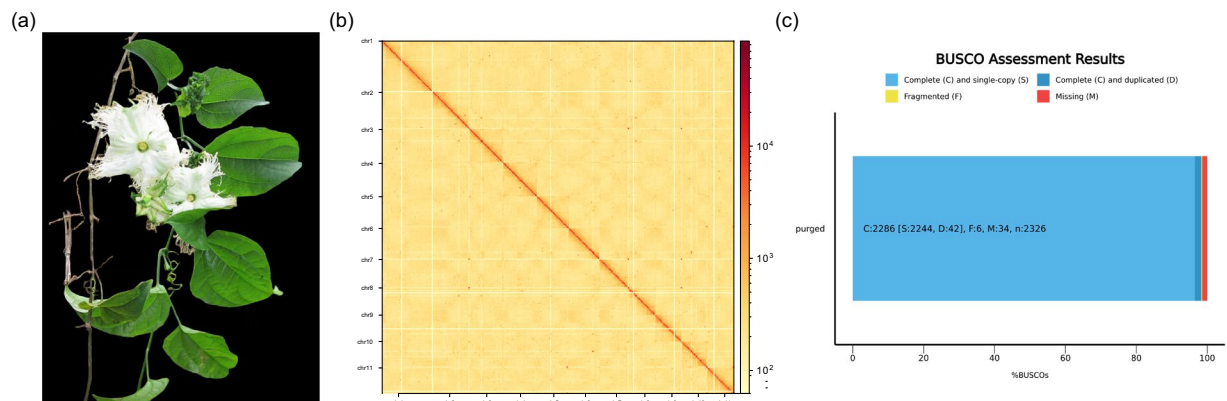


Fig. 1 Genome assembly of *Trichosanthes truncata* C. B. Clarke. **(a)** Photo of *T. truncata* C. B. Clarke. **(b)** Hi-C heatmap showing chromosomal interactions under a resolution of 1 Mb. Only sequences anchored on chromosomes are shown. The abbreviations chr1–11 represent the 11 chromosomes, and the color bar represents the log₂ value of interaction counts. **(c)** Completeness of the expected gene space of the genome assembly estimated with BUSCO.

sent to Shenzhen BGI Genomic for sequencing. Total genomic DNA was extracted from the frozen leaves using cetyltrimethylammonium bromide (CTAB) buffer (incubation for 60 min at 65 °C) and was purified using phenol/chloroform/isopentyl (25:24:1), isopropyl alcohol, and ethanol precipitation. Purified DNA was resuspended in Tris–EDTA buffer for sequencing.

Library construction and genome sequencing. The extracted DNA samples were subjected to Megaruptor shearing, resulting in fragment sizes within the target range of 15–20 kilobases (Kb). Subsequently, the DNA underwent damage repair, end repair, ligation with known adapters, enzymatic digestion, and size selection using the BluePippin system (Sage Science, Beverly, MA, USA), ultimately yielding dumbbell-shaped libraries. After quality control checks, these libraries were utilized for PacBio HiFi sequencing.

Two short-read sequencing libraries with insert sizes of 270 bp and 500 bp were constructed from the resulting high-quality DNA. DNA was subjected to fragmentation (Covaris, Woburn, MA, USA) and end repair, followed by adaptor ligation, which enabled the formation of circular DNA molecules and subsequent rolling-circle amplification to produce DNA nanoballs (DNBs). To prepare a Hi-C library, cells were treated with formaldehyde to cross-link DNA–protein or protein–protein complexes, which were then subjected to fragmentation, end repair, purification, and adaptor ligation. The short-read sequencing libraries and the Hi-C library were sequenced using the DNBSEQ platform (MGI, Shenzhen, China) in paired-end mode.

Genome size estimation. The genome size of *T. truncata* was estimated based on k-mer counting of sequencing data using the k-mer analysis toolkit v2.4.2⁸, followed by genome assessment, computation of genome size, and heterozygosity estimation using genomescope2⁹.

Genome assembly and assessment. We assembled PacBio HiFi reads using hifiasm v0.16.1 and validated the correctness of the k-mer peak positions¹⁰. Subsequently, we employed purge_dups v1.2.6 (https://github.com/dfguan/purge_dups) to further refine the p_ctg.gfa results by de-duplication and eliminating haplotigs with excessively low or high coverage generated during the assembly process. The Arima mapping pipeline (https://github.com/ArimaGenomics/mapping_pipeline) was used to align the Hi-C data with the assembled genome contig sequences. Scaffolding was performed using SALSA2¹¹, followed by manual curation and refinement of the scaffolding using Juicebox (JBAT) (<https://github.com/aidenlab/Juicebox>).

The completeness and duplication levels of genome assembly were assessed throughout the process using BUSCO v5.2.2¹². The Extensive de novo TE Annotator (EDTA) pipeline v2.0.1¹³ was employed to annotate TEs, which were then used to softmask the genome assembly with RepeatMasker v4.1.2¹⁴. Annotation of protein-coding genes was performed using a combination of homology-based searches and de novo predictions. First, protein sequences of 10 known species from the Cucurbitaceae family including *T. anguina*, *Citrullus lanatus*, *Luffa aegyptiaca*, *Cucurbita moschata*, *Siraitia grosvenorii*, *Momordica charantia*, *Gynostemma pentaphyllum*, *Cucumis sativus*, *Lagenaria siceraria*, and *Benincasa hispida*, were aligned to the *T. truncata* genome using diamond v2.0.15¹⁵. Subsequently, de novo gene prediction was performed using Funannotate pipeline v1.8.7 (<https://github.com/nextgenusfs/funannotate>) including AUGUSTUS¹⁶, GlimmerHMM v3.02¹⁷, and SNAP v2006-07-28¹⁸. Based on protein sequences from the 10 related species and the predicted gene models, the annotation data were combined to generate final consensus gene models using EVIDENCEModeler v1.1.1¹⁹, which were then assessed using BUSCO with the “embryophyta_odb10” ortholog set¹². The tRNAs were predicted using tRNAscan-SE v2.0²⁰, and ribosomal RNAs were identified using RNAmmer²¹. PANTHER families were predicted using the program InterProScan v5.41–78.0²². Gene functional annotation was performed using the eggNOG-mapper software v2²³ and the evolutionary genealogy of genes (eggNOG) database v5.0²⁴.

Estimated genome size	557.50 Mb
Heterozygosity	1.22%
Assembly	
Total length (Mb)	637.41
GC content (%)	33.82
Coting N50 (Mb)	56.40
Contig number	154
Scaffold N50 (Mb)	57.24
Scaffold number	83
Pseudochromosome number	11
BUSCO completeness score (%)	98.30
Annotation	
Total length of repeats (Mb)	97.21
Number of protein-coding genes	36172
Mean transcript length (bp)	1015.65
Mean coding sequence length (bp)	2999.48
Mean exon length	227.57

Table 1. Statistics of the Assembly and Annotation of the *Trichosanthes Kirilowii* Maxim. Genome.

Comparative genome analyses. Gene family clustering analysis included 11 species from the Cucurbitaceae family, as well as rice and papaya: *T. truncata*, *T. anguina*, *Citrullus lanatus*, *Luffa aegyptiaca*, *Cucurbita moschata*, *Siraitia grosvenorii*, *Momordica charantia*, *Gynostemma pentaphyllum*, *Cucumis sativus*, *Lagenaria siceraria*, *Benincasa hispida*, *Oryza sativa*, and *Carica papaya*. For each gene, only the transcript with the longest coding region was retained and genes encoding proteins with fewer than 50 amino acids were excluded. After filtering the gene sets for each species, the remaining protein sequences were aligned using Diamond¹⁵, and low-quality alignments were filtered using a default similarity threshold of 95%. Subsequently, the alignment results were clustered using the proteinortho software to generate protein orthologs²⁵.

To detect whole-genome duplication (WGD) events within the *T. truncata* genome, we employed the MCScanX to delineate syntenic blocks and subsequently computed the 4DTv for all gene pairs in each syntenic segment²⁶.

Gene families with abnormally high or low gene count within a species were filtered out, and expansion and contraction analyses of gene families were performed using CAFE5^{27,28}. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of the expanded and contracted gene families was performed with clusterProfiler v4.8.3²⁹.

Data Records

The assembly is available at Genbank under the accession number GCA_033996785.1³⁰. The raw genome sequencing data (PacBio and DNBSEQ short reads) have been deposited to NCBI database under the SRP accession number SRP465770³¹ and SRR29025305³².

Technical Validation

Genome sequencing and assembly. PacBio HiFi sequencing and Hi-C techniques were used to generate a chromosome-scale assembly of *T. truncata* C. B. Clarke. The total length of the final assembly was 637.41 Mb, with a GC content of 33.82% (Table 1), which was close to the genome size estimated by 25-mer analysis (genome size of 557.50 Mb and heterozygosity of 1.217%). The numbers of contigs and scaffolds were 154 and 83, respectively, and contig N50 and scaffold N50 were approximately 56.40 Mb and 57.24 Mb, respectively (Table 1). Eleven chromosomes were generated by concatenating the contigs based on the Hi-C reads (Table 1), which is consistent with the findings of previous studies based on karyotype analysis³³. The interaction signal intensity around the diagonal of the genome-wide Hi-C heatmap was stronger than that of the off-diagonal signals (Fig. 1b), indicating the high quality of the chromosome-level genome assembly.

Genome annotation. The total length of the repeats was 97.21 Mbp, accounting for 15.25% of the reference genome sequence (Table 1). A comprehensive approach was employed to perform genome annotation in *T. truncata*, combining homology-based searches and multiple prediction algorithms (AUGUSTUS, GlimmerHMM, SNAP). We identified 36,172 protein-coding genes across the genome assembly, with an average transcript length of 1,015.65 bp (Table 1). The average lengths of the coding sequence and exons were 2,999.48 bp and 227.57 bp, respectively (Table 1). BUSCO evaluation revealed a completeness score of approximately 98.30% based on the genes in the “Embryophyta_odb10” ortholog set (Fig. 1c and Table 1), indicating a high degree of completeness³⁰. While these methods provided a robust initial annotation, integrating RNA-seq data in future studies will be pivotal for refining gene annotations and improving accuracy, thereby enhancing our understanding of gene function and genomic dynamics in *T. truncata*.

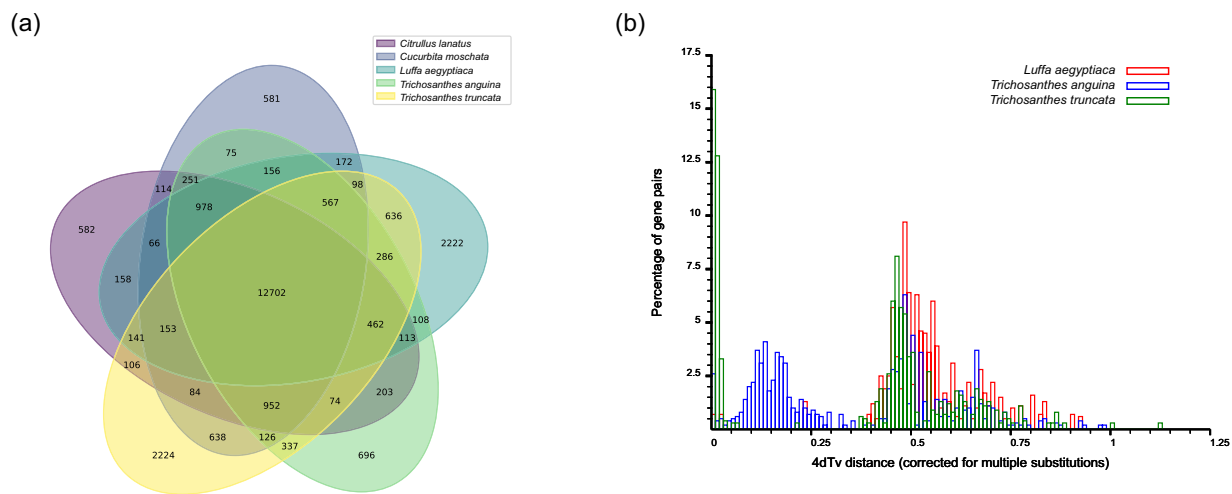


Fig. 2 Gene family and evolution analysis. **(a)** Venn diagram representing a comparison of gene clusters among *T. truncata*, *C. lanatus*, *C. moschata*, *L. aegyptiaca*, and *T. anguina* L. **(b)** Detection of whole-genome duplication events of *T. truncata* by 4-fold degenerate synonymous sites (4dTv) comparisons.

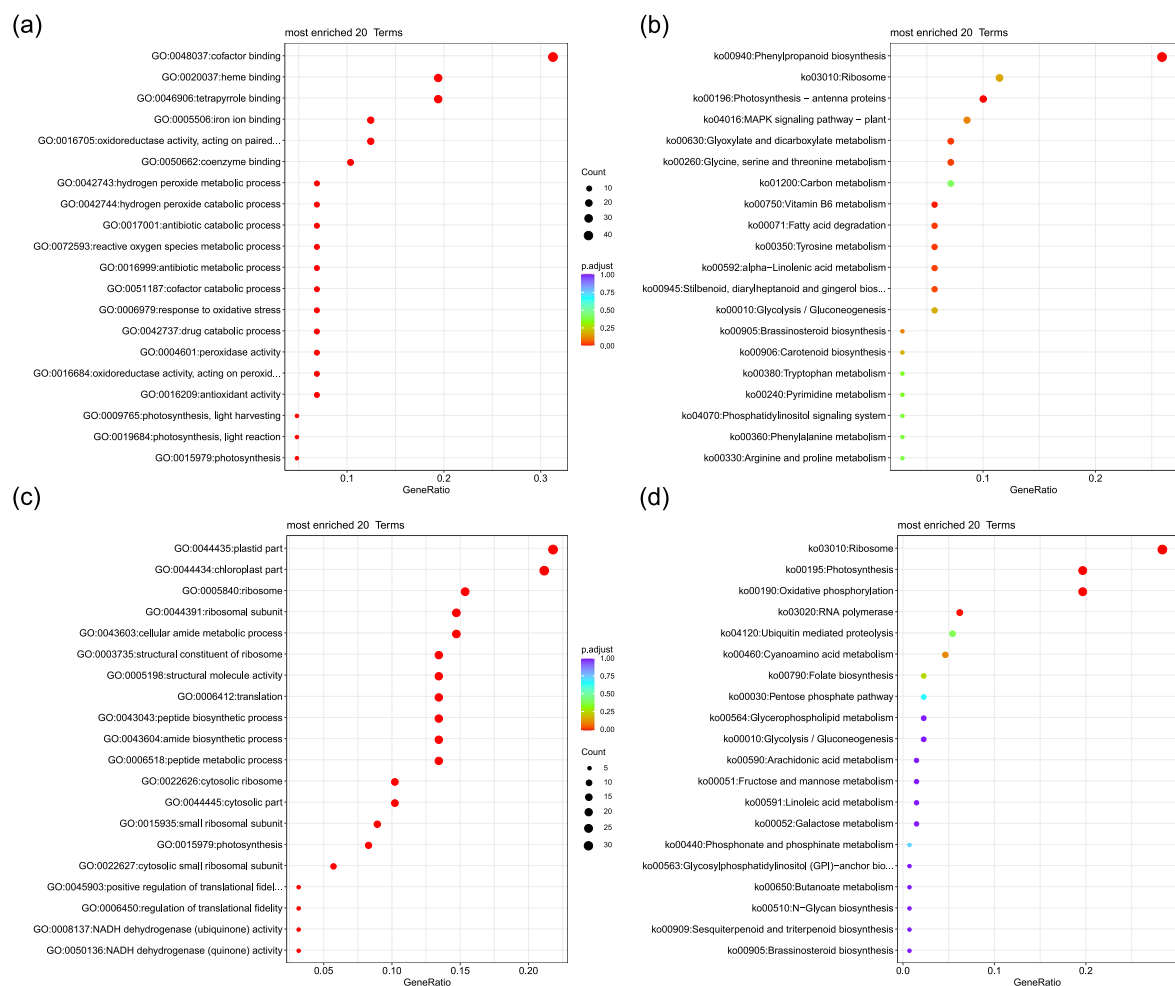


Fig. 3 GO **(a,c)** and KEGG pathway **(b,d)** enrichment analyses of expanded **(a,b)** and contracted **(c,d)** gene families. The GO terms and KEGG pathways with an FDR-adjusted *p*-value < 0.05 were defined as statistically significant.

Gene family analysis. Gene family cluster analysis of complete gene sets of *T. truncata* and 12 other sequenced plant genomes, including *T. anguina*, *C. lanatus*, *L. aegyptiaca*, *C. moschata*, *S. grosvenorii*, *M. charantia*, *G. pentaphyllum*, *C. sativus*, *L. siceraria*, *B. hispida*, *O. sativa*, and *C. papaya* was performed. All gene families were grouped into 37,061 gene clusters, with 1,060 single-copy orthologs. Further comparison of *T. truncata* with four closely related species (*C. lanatus*, *C. moschata*, *L. aegyptiaca*, and *T. anguina*) revealed 12,702 gene clusters shared by the five species. In addition, we found 2,224 gene families that were unique to the *T. truncata* genome when compared with the other four genomes (Fig. 2a).

Comparative genome analysis. Fourfold degenerate synonymous site (4DTv) comparisons on gene synteny blocks revealed a recent *T. truncata*-specific whole-genome duplication (WGD) event (Fig. 2b). An expanded analysis of gene families was performed using 12 other sequenced plant genomes, which indicated that 158 gene families expanded in *T. truncata*, whereas 1140 gene families contracted, suggesting that more *T. truncata* gene families experienced contractions than expansions during adaptive evolution.

The expanded gene families were found to be primarily enriched in GO functions related to photosynthesis, including light reaction, light harvesting, antioxidant activity, oxidoreductase activity (acting on peroxide as an acceptor), oxidoreductase activity (acting on paired donors with the incorporation or reduction of molecular oxygen), peroxidase activity, hydrogen peroxide catabolic process, reactive oxygen species metabolic process, cofactor catabolic process, and antibiotic catabolic process (Fig. 3a). In terms of KEGG pathways, enrichment was observed in phenylpropanoid biosynthesis, Photosynthesis - antenna proteins, vitamin B6 metabolism, fatty acid degradation, tyrosine metabolism, glyoxylate and dicarboxylate metabolism, glycine, serine and threonine metabolism, alpha-linolenic acid metabolism, and stilbenoid, diarylheptanoid, and gingerol biosynthesis (Fig. 3b). On the other hand, the contracted gene families were mainly enriched in GO functions associated with chloroplast components, NADH dehydrogenase (quinone) activity, NADH dehydrogenase (ubiquinone) activity, positive regulation of translational fidelity, photosynthesis, amide biosynthetic process, and peptide biosynthetic process (Fig. 3c). In terms of KEGG pathways, enrichment was observed in photosynthesis, oxidative phosphorylation, ribosome, RNA polymerase, and cyanoamino acid metabolism (Fig. 3d).

Code availability

No specific code was used in this study. The data analyses used standard bioinformatic tools, with parameters being clearly described in Methods. If specific parameters were not provided for the software, default settings recommended by the developer were utilized.

Received: 7 December 2023; Accepted: 5 July 2024;

Published online: 12 August 2024

References

- Kumar, N., Singh, S., Manvi, S. & Gupta, R. Trichosanthes dioica Roxb.: An overview. *Pharmacogn Rev.* **6**, 61–67 (2012).
- Fang, E. F., Ng, T. B., Shaw, P. C. & Wong, R. N. Recent progress in medicinal investigations on trichosanthin and other ribosome inactivating proteins from the plant genus Trichosanthes. *Curr. Med. Chem.* **18**, 4410–4417 (2011).
- Changsan, Z., Rong, C., Suping, L. & Jiacheng, A. Study on seed examination and germination characteristics of Trichosanthes truncata. *Journal of Jiangsu Agricultural Sciences* **42**, 270–272 (2014).
- Meilan, H. & Weijian, B. Comparative Pharmacological Effects of Trichosanthes truncata and Trichosanthes Species. *Guiding Journal of Traditional Chinese Medicine and Pharmacy* (2000).
- de Boer, H. J. & Thulin, M. Synopsis of Trichosanthes (Cucurbitaceae) based on recent molecular phylogenetic data. *PhytoKeys*, 23–33 (2012).
- Ma, L. *et al.* The genome and transcriptome analysis of snake gourd provide insights into its evolution and fruit development and ripening. *Hortic Res.* **7**, 199 (2020).
- Song, P., Xu, B., Liu, Z., Cheng, Y. & Chao, Z. The Difference of Volatile Compounds in Female and Male Buds of Trichosanthes anguina L. Based on HS-SPME-GC-MS and Multivariate Statistical Analysis. *Molecules* **27** (2022).
- Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J. & Clavijo, B. J. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* **33**, 574–576 (2017).
- Vurtture, G. W. *et al.* GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
- Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
- Ghurye, J. *et al.* Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput. Biol.* **15**, e1007273 (2019).
- Manni, M., Berkeley, M. R., Seppely, M., Simao, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).
- Ou, S., Su, W., Liao, Y., Chougule, K. & Hufford, M. B. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
- Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2020).
- Buchfink, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
- Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**, 215–225 (2003).
- Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**, 4636–41 (1999).
- Korf, I. Gene finding in novel genomes. *BMC Bioinform.* **5**, 59 (2004).
- Haas, B. J., Salzberg, S. L., Zhu, W. & Pertea, M. Automated eukaryotic gene structure annotation using EvidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
- Chan, P. P. & Lowe, T. M. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. *Methods Mol. Biol.* **1962**, 1–14 (2019).
- Lagesen, K. *et al.* RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108 (2007).
- Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

23. Huerta-Cepas, J. *et al.* Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
24. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, 309–314 (2019).
25. Lechner, M. *et al.* Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics* **12**, 124 (2011).
26. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
27. De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269–1271 (2006).
28. Mendes, F. K., Vanderpool, D., Fulton, B. & Hahn, M. W. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* **36**, 5516–5518 (2021).
29. Wu, T. Z. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* **2**, 100141 (2021).
30. Guangxi Botanical Garden of Medicinal Plants. *GenBank* https://identifiers.org/insdc.gca:GCA_033996785.1 (2023).
31. *NCBI Sequence Read Archive* <https://identifiers.org/ncbi/insdc.sra:SRP465770> (2023).
32. *NCBI Sequence Read Archive* <https://identifiers.org/ncbi/insdc.sra:SRR29025305> (2024).
33. Huang, L. Q. *et al.* Phylogenetic studies of the genus *Trichosanthes* L. *J. Jiangxi Coll. Trad. Chin. Med.* **11**, 75–78 (1999).

Acknowledgements

Thanks are due to Dr Jianhua Miao for assistance with the experiments. This work was supported by the Guangxi Innovation-Driven Development Project (GuiKe AA18242040), Key Project at Central Government Level (2060302), the earmarked fund for Modern Agro-industry Technology Research System (CARS-21).

Author contributions

Y.H., K.H.W. and W.W.L. conceived the project and designed the experiments. X.M.W., Z.N.C. and F.W. conducted the experiments. Y.H., Y.D.P., B.Y.H. and L.D. analyzed the results. Y.H., K.H.W. and W.W.L. wrote the paper. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to K.W. or W.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024