scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A chromosome-level genome assembly of the forestry pest *Coronaproctus castanopsis*

Yi-Xin Huang[1,2,3], Xiu-Shuang Zhu[3], Xiao-Nan Chen[2], Xin-Yi Zheng[4], Bao-Shan Su[3], Xiao-Yu Shi[1], Xu Wang[1,5], San-An Wu[4], Hao-Yuan Hu[3], Jian-Ping Yu[2]✉, Yan-Zhou Zhang[1]✉ & Chao-Dong Zhu[1]

As an important forestry pest, *Coronaproctus castanopsis* (Monophlebidae) has caused serious damage to the globally valuable Gutianshan ecosystem, China. In this study, we assembled the first chromosome-level genome of the female specimen of *C. castanopsis* by merging BGI reads, HiFi long reads and Hi-C data. The assembled genome size is 700.81 Mb, with a scaffold N50 size of 273.84 Mb and a contig N50 size of 12.37 Mb. Hi-C scaffolding assigned 98.32% (689.03 Mb) of *C. Castanopsis* genome to three chromosomes. The BUSCO analysis (n = 1,367) showed a completeness of 91.2%, comprising 89.2% of single-copy BUSCOs and 2.0% of multicopy BUSCOs. The mapping ratio of BGI, second-generation RNA, third-generation RNA and HiFi reads are 97.84%, 96.15%, 97.96%, and 99.33%, respectively. We also identified 64.97% (455.3 Mb) repetitive elements, 1,373 non-coding RNAs and 10,542 protein-coding genes. This study assembled a high-quality genome of *C. castanopsis*, which accumulated valuable molecular data for scale insects.

Background & Summary

Scale insects are highly adaptable to the surrounding environment and are widespread throughout the world, with more than 8520 species in 56 families (36 extant families and 20 extinct families) recorded to date. With the exception of a few resource species that can be applied to the chemical industry, such as *Ericerus pela*[1], *Dactylopius coccus*[2] and *Laccifer lacca*[3], most scale insect are important agroforestry pests.

Coronaproctus castanopsis Li, Xu & Wu, 2023, was firstly discovered in Gutianshan National Nature Reserve, China[4]. Globally unique and undisturbed low-altitude subtropical evergreen broadleaf forests can be found in the Gutianshan Reserve. The field survey revealed that *C. castanopsis* are oligophagous and some of its main host plants are *Castanopsis eyrei*, *Castanopsis carlesii*, and *Castanopsis fargesii* (Fagaceae). These three species of trees are the primary constituents of the forest ecosystem in the reserve, and the scale insects mostly reside on the tree crowns, which are often difficult to observe. As a result, *C. castanopsis* has caused serious damage to the forest ecosystems of the Gutianshan Reserve.

The difficulty of high-quality scale insect genome assembly lies in its high degree of heterozygosity and a large number of repetitive sequences. There are only 13 coccoid genomes in the GenBank database, of which four species, mealybug - *Balanococcus diminutus, Phenacoccus solenopsis*, *Planococcus citri* and giant mealybug - *Icerya purchasi*, have been assembled into the chromosome-level genome. The limited availability of genomic data hindered our research on this group. Therefore, we constructed a chromosome-level *C. castanopsis* genome using a combination of BGI short reads, hifi long reads, and Hi-C data. We also annotated the genome for repetitive elements, protein-coding genes and non-coding RNAs, and performed phylogenetic and evolutionary analysis of the gene family. Our results contribute to the genome database of Coccomorpha and offer substantial support for a deeper understandingof *C. castanopsis* and future studies into scale insects.

[1]Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing, 100101, China. [2]Qianjiangyuan National Park, Kaihua, Zhejiang, 324300, China. [3]Collaborative Innovation Center of Recovery and Reconstruction of Degraded Ecosystem in Wanjiang Basin Co-founded by Anhui Province and Ministry of Education, School of Ecology and Environment, Anhui Normal University, Wuhu, Anhui, 241000, China. [4]Beijing Forestry University, Beijing, 100083, China. [5]Anhui Provincial Key Laboratory of the Conservation and Exploitation of Biological Resources, College of Life Sciences, Anhui Normal University, Wuhu, Anhui, 241000, China. ✉e-mail: 1125142830@qq.com; zhangyz@ioz.ac.cn

Fig. 1 Ecological photo of a female adult *C. castanopsis* (photographed by Xiu-Shuang Zhu).

Genomic libraries	Clean data (Gb)	Mean length (bp)	N50 (kb)	Sequencing coverage (X)
BGI	67.19	150	—	155.64
HiFi	36.75	16,184.60	16.18	52.50
Hi-C	58.78	150	—	83.97
RNA-sr	7.87	150	—	—
RNA-ONT	13.40	926.87	1.13	—

Table 1. Sequencing data statistics for genome assembly.

Methods

Samples collection and sequencing. Adult female specimens of *C. castanopsis* (Fig. 1) were collected in May 2022 at Gutianshan National Nature Reserve (29.265° N, 118.101° E), Quzhou city, Zhejiang Province, China. Fresh samples were immediately placed in liquid nitrogen after collection and then stored at −80 °C for further use. To reduce contamination from gut microbes, we removed the metasoma of the samples and sent them to Berry Genomics Corporation (Beijing, China) for genome sequencing. The number of individuals used for genome survey, PacBio, Hi-C, and transcriptome sequencing was 10, 3, 5 and 5, respectively. Adult female specimens of *C. castanopsis* were used for transcriptome sequencing.

Genomic DNA, second-generation RNA and third-generation full-length RNA were extracted using the CTAB method[5], the TRIzol TM Reagent Kit, and the RNA prep Pure Plant Plus Kit, respectively. The second-generation genome sequencing was completed on the Beijing Genomics Institute platform, and BGISEQ-500 library was constructed using the Agencourt AMPure XP-Medium Kit (insert size: 350 bp). PacBio HiFi sequencing was performed on the PacBio Sequel IIe platform, and the PacBio HiFi 15 K library was constructed using the SMRTbell® Express Template Prep Kit 2.0. Third-generation RNA sequencing (Oxford Nanopore Technologies (ONT) Oxford, UK) was performed on the Oxford Nanopore PromethION platform, and the ONT PromethION library was constructed using the SQK-PCS109 and SQKPBK004 kit. Both second-generation RNA (RNA-sr) and Hi-C library were performed on the Illumina NovaSeq. 6000 platform with 150-bp paired-end reads. We totally sequenced 183.99 Gb clean reads, including 36.75 Gb (53x) PacBio reads (N50 16.18 kb), 67.19 Gb (156x) BGI reads, 58.78 Gb (84x) Hi-C reads, 7.87 Gb second-generation RNA reads, and 13.40 Gb ONT RNA reads (Table 1).

Genome assembly. High-quality HiFi reads (Q20 base quality) were generated by pbccs v6.4.0. Hifiasm v0.16.1[6] was used for the first round of assembly with a parameter setting of " -l 2". The Hifiasm assembly only retained contig sequences with a sequencing depth of more than 10X to avoid possible errors or contamination. Minimap2 v2.24[7] was used to paste the second-generation data back to the Hifiasm assembly, and SAMtools v1.10[8] was used to convert the data format sam to bam. We also used NextPolish v1.4.0[9] to perform short-read

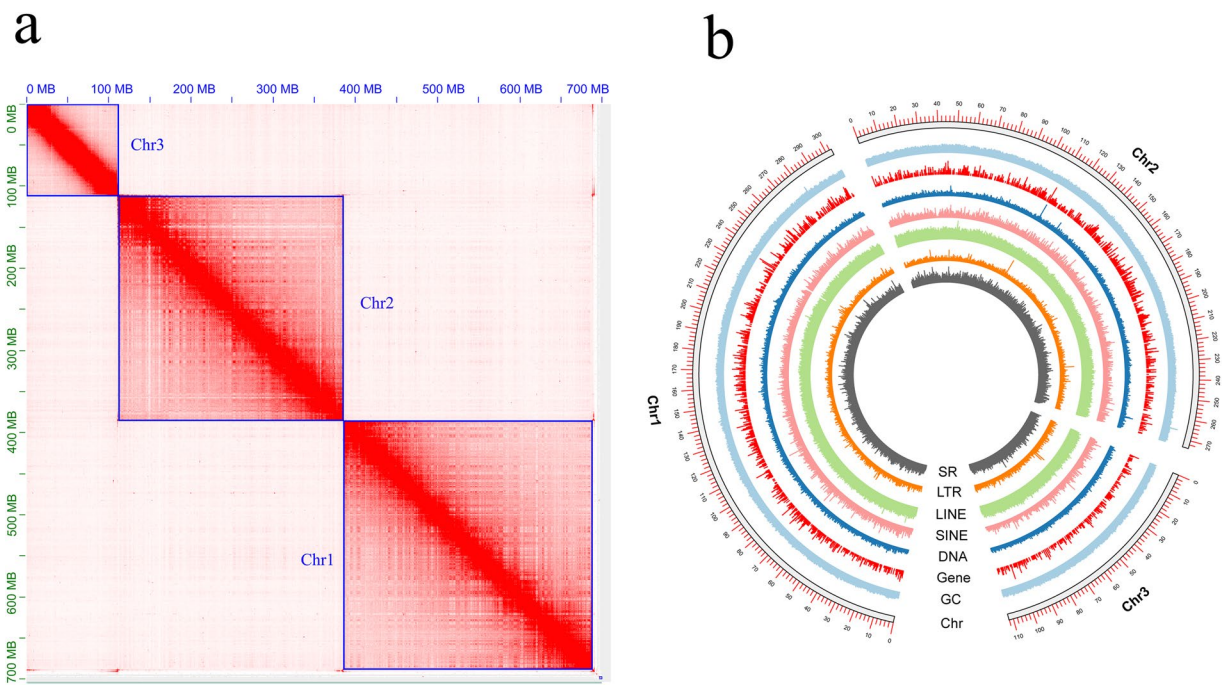

Fig. 2 Genomic heatmap and features. (**a**) genome-scale chromosome heatmap of *C. castanopsis*, with individual chromosome outlined in blue. (**b**) circos plot with a window size of 100 Kbp. Each circle from inside to outside represents simple repeats, LTR, LINE, SINE, DNA, gene density, GC content and chromosome length.

and long-read polishing to improve assembly accuracy. The Hi-C data and the 3D-DNA v180922[10] process were used for chromosome mounting and assembly of contigs. After using Juicer v1.6.2[11] to perform quality control on Hi-C data, we then performed two rounds of splicing using the default parameters of 3D-DNA v180922. Manual error correction was performed using Juicebox v1.11.08, and the sequencing depth of each pseudochromosome was evaluated by bamtocov v.2.7.0[12]. Genomic integrity was assessed by BUSCO v5.2.2[13] based on the insecta_odb10 database (n = 1,367). Next, we used the postback tool Minimap2 to test the utilization of the original data and the integrity of the assembly, and the postback rate was counted by SAMtools v1.10. After polishing and correction, the final assembled genome size of *C. castanopsis* was 700.81 Mb, including 53 scaffolds and 161 contigs, with the scaffold/contig N50 size of 273.84/12.37 Mb and a GC content of 31.58% (Fig. 2a,b). In addition, Hi-C scaffolding assigned 98.32% (689.03 Mb) of *C. Castanopsis* genome to three pseudo-chromosomes (Fig. 2a). The BGI, second-generation RNA, third-generation RNA, and HiFi data reply rates were 97.84%, 96.15%, 97.96%, and 99.33%, respectively. The BUSCO analysis (n = 1,367) showed a completeness of 91.2%, comprising 89.2% of single-copy BUSCOs and 2.0% of multicopy BUSCOs. The above indicators showed that the assembly has reached a high level in terms of both continuity and integrity. We note that the values in the article may differ slightly in the final version of this assembly, where ~0.01% of the bases were removed or masked by the NCBI contamination screening program. In general, the genome of *C. castanopsis* has been assembled to a high degree of completeness.

Genome annotation. Using RepeatMasker v4.1.2p1 (http://www.repeatmasker.org), we identified the repetitive regions of the genome against the final repetitive sequence reference database. The final repetitive sequence reference database included de novo repeat library, Dfam 3.5[14] and RepBase-20181026[15]. The de novo repeat library was constructed using RepeatModeler v2.0.3[16] and the '-LTRStruct' search process. The results showed that the *C. castanopsis* genome contains about 64.97% (455.3 Mb) of repetitive elements, including LINEs (39.60%), unclassifed elements (13.15%), DNA transposons (6.33%), LTR elements (1.39%), simple repeats (2.68%), and other elements (Table S1).

To predict and identify the protein-coding gene structure, we used MAKER v3.01.03[17] to integrate three types of strategies (ab initio prediction, transcript sequence alignment and homologous proteins comparison). Input files for MAKER ab initio were obtained by using BRAKER v2.1.6[18] and GeMoMa v1.8[19] and integrating both transcriptomic and protein evidence. Transcriptome alignments were generated by using HISAT2 v2.2.0[20]. Two predictors, Augustus v3.3.4[21] and GeneMark-ES/ET/EP 4.68_3.60_lic[22], were automatically trained by BRAKER based on reference proteins mined from the OrthoDB10 v1 database[23] and transcriptome data. Using information on protein homology and intron location, GeMoMa was used to predict genes with the parameter of "GeMoMa.c = 0.4GeMoMa.p = 10" and the protein sequences of five related species (*Tribolium castaneum*, *Coccinella septempunctata*, *Apis mellifera*, *Chrysoperla carnea* and *Drosophila melanogaster*). Reference assembly (–mix) based on second and third-generation transcriptomes was performed using StringTie v2.1.6[24], and RNA sequences alignments were generated by HISAT2. Besides, predictions were made in GeMoMa via homology comparison with the protein sequences of the five species above. In total, the MAKER process identified 10,542 protein-coding genes with an average gene length of 19,827.3 bp. The average number of exons (mean length:

294.3 bp), introns (mean length: 2629.6 bp) and CDS (mean length: 208 bp) in each gene was 7.8, 6.8 and 7.5, respectively. The predicted protein gene sequences assessed for BUSCO completeness were 91.2% (n: 1367), including 78.8% single-copy, 12.4% duplicated, 0.7% fragmented and 8.1% missing BUSCOs.

Using the high-sensitivity mode (–very-sensitive -e 1e-5) in Diamond v2.0.11.149[25], we searched the UniProtKB database for protein-coding gene function annotation. In addition, in order to annotate Gene Ontology (GO) and (KEGG, Reactome) pathways and identify protein domains, we searched Pfam[26], SMART[27],Superfamily[28] and CDD[29] databases using InterProScan 5.53–87.0[30], and we also searched the eggNOG v5.0[31] database using eggNOG-mapper v2.1.5[32]. Finally, Genes with 8363 GO terms, 4217 KEGG pathways, 2474 Enzyme Codes, 7982 Reactome pathways, and 9323 COG categories were identified by combining the eggNOG and InterProScan annotation results (Table S2).

The annotations of rRNA, snRNA and miRNA were compared with the Rfam database using Infernal v1.1.4[33]. Prediction of tRNA sequences was performed using tRNAscan-SE v2.0.9[34], with low confidence tRNAs filtered by the 'EukHighConfidenceFilter' script. We totally identifed 1373 ncRNAs in the genome of *C. castanopsis*, including 265 ribosomal RNAs, 52 microRNAs, 22 small RNAs, 40 long non-coding RNA, 515 small nuclear RNAs, 153 transfer RNAs, and 326 other ncRNAs (Table S3).

Data Records

The raw sequencing data and genome assembly of *Coronaproctus castanopsis* have been submitted to the National Center for Biotechnology Information (NCBI) and the China National GeneBank DataBase (CNGBdb). The Hi-C, PacBio, RNA-ONT, RNA-sr and BGI data are accessible via accession numbers SRR 26067557-SRR26067561[35–39]. The BGI, RNA-sr, RNA-ONT, PacBio and Hi-C data are accessible via accession numbers CNX0846626-CNX0846630[40–44]. The assembled genome is accessible via accession number GCA_032883995.1[45].

Technical Validation

The assessment of the quality of the genome assembly has been a two-step process. Initially, we assessed the completeness of the assembly using BUSCO v5.2.2 based on the insecta_odb10 database (n = 1,367). The final genome assembly displayed a BUSCO completeness of 91.2%, comprising of 1219 (89.2%) single-copy BUSCOs, 27 (2.0%) duplicated BUSCOs, 33 (2.4%) fragmented BUSCOs, and 87 (6.4%) missing BUSCOs. We then calculated the mapping rate to measure the accuracy of the assembly. The BGI, second-generation RNA, third-generation RNA, and Hifi data reply rates were 97.84%, 96.15%, 97.96%, and 99.33%, respectively. Overall, these assessments reflect the high quality of the genomic assembly.

Code availability

No specific script was used in this work. All commands and pipelines used in data processing were executed according to the manual and protocols of the corresponding bioinformatic softwares.

Received: 17 October 2023; Accepted: 22 January 2024;
Published online: 17 February 2024

References

1. Yang, P. *et al*. Genome sequence of the Chinese white wax scale insect *Ericerus pela*: the first draft genome for the Coccidae family of scale insects. *Gigascience*. **8**, 1–8 (2019).
2. Campana, M. G., Robles García, N. M. & Tuross, N. America's red gold: multiple lineages of cultivated cochineal in mexico. *Ecol Evol*. **5**, 607–617 (2015).
3. Patel, A. R. & Dewettinck, K. Comparative evaluation of structured oil systems: Shellac oleogel, HPMC oleogel, and HIPE gel. *Eur J Lipid Sci Tech*. **117**, 1772–1781 (2015).
4. Li, J., Xu, H. & Wu, S. A. A new genus and species of giant mealybugs (Hemiptera: Coccomorpha: Monophlebidae) from eastern China. *Zootaxa*. **5254**, 434–442 (2023).
5. Shahjahan, R. M., Hughes, K. J., Leopold, R. A. & Devault, J. D. Lower incubation temperature increases yield of insect genomic DNA isolated by the CTAB method. *Biotechniques*. **19**, 332–334 (1995).
6. 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).
7. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics*. **37**, 4572–4574 (2021).
8. Li, H. *et al*. The Sequence Alignment/Map Format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).
9. Hu, J., Fan, J., Sun, Z. Y., Liu, S. L. & Berger, B. NextPolish: a fast and efficient genome polishing tool for long read assembly. *Bioinformatics*. **36**, 2253–2255 (2020).
10. Dudchenko, O. *et al*. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science*. **356**, 92–95 (2017).
11. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst*. **3**, 95–98 (2016).
12. Birolo, G. & Telatin, A. BamToCov: an efficient toolkit for sequence coverage calculations. *Bioinformatics*. **38**, 2617–2618 (2022).
13. Manni, M., Berkeley, M. R., Seppey, M., Simão, 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).
14. Hubley, R. *et al*. The Dfam database of repetitive DNA families. *Nucleic Acids Res*. **44**, D81–D89 (2016).
15. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*. **6**, 1–6 (2015).
16. Flynn, J. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA* **117**, 9451–9457 (2020).
17. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*. **12**, 491 (2011).
18. Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1: unsupervised RNA-Seq-Based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*. **32**, 767–769 (2016).

19. Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S. O. & Grau, J. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinformatics*. **19**, 189 (2018).
20. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: A fast spliced aligner with low memory requirements. *Nat Methods*. **12**, 357–360 (2015).
21. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, W309–W312 (2004).
22. Brůna, T., Lomsadze, A. & Borodovsky, M. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genom Bioinform.* **2**, 1–14 (2020).
23. Kriventseva, E. V. *et al*. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* **47**, D807–D811 (2019).
24. Kovaka, S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).
25. Buchfink, B. *et al*. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*. **18**, 366–368 (2021).
26. El-Gebali, S. *et al*. The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
27. Letunic, I. & Bork, P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.* **46**, D493–D496 (2018).
28. Wilson, D. *et al*. SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.* **37**, D380–D386 (2009).
29. Marchler-Bauer, A. *et al*. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* **45**, D200–D203 (2017).
30. Finn, R. D. *et al*. InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.* **45**, D190–D199 (2017).
31. 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 Research*. **47**, D309–D314 (2019).
32. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*. **38**, 5825–5829 (2021).
33. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*. **29**, 2933–2935 (2013).
34. Chan, P. P. & Lowe, T. M. tRNAscan-SE: searching for tRNA genes in genomic sequences. *Methods Mol Biol.* **1962**, 1–14 (2019).
35. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR26067557 (2023).
36. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR26067558 (2023).
37. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR26067559 (2023).
38. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR26067560 (2023).
39. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR26067561 (2023).
40. *CNGBdb Sequence Read Archive* https://db.cngb.org/search/experiment/CNX0846626/ (2023).
41. *CNGBdb Sequence Read Archive* https://db.cngb.org/search/experiment/CNX0846627/ (2023).
42. *CNGBdb Sequence Read Archive* https://db.cngb.org/search/experiment/CNX0846628/ (2023).
43. *CNGBdb Sequence Read Archive* https://db.cngb.org/search/experiment/CNX0846629/ (2023).
44. *CNGBdb Sequence Read Archive* https://db.cngb.org/search/experiment/CNX0846630/ (2023).
45. *NCBI Assembly* https://identifiers.org/ncbi/insdc.gca:GCA_032883995.1 (2023).

Author contributions

Y.X.H., X.N.C., S.A.W., H.Y.H., J.P.Y., Y.Z.Z. and C.D.Z. contributed to the research design. Y.X.H., X.S.Z., B.S.S. and X.W. collected the samples. Y.X.H., X.S.Z., X.Y.Z., B.S.S., X.Y.S. and X.W. analyzed the data. Y.X.H., X.S.Z., X.N.C., X.Y.Z., X.Y.S., S.A.W., H.Y.H., J.P.Y., X.Y.Z. and C.D.Z. wrote the draf manuscript and revised the manuscript. All co-authors contributed to this manuscript and approved it.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-024-03016-6.

Correspondence and requests for materials should be addressed to J.-P.Y. or Y.-Z.Z.