scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly and annotation of the predatory stink bug *Eocanthecona furcellata*

De-Long Guan[1,2], Ying-can Qin[1], Shi-Hao Zhang[1,2], Yue Qin[1,2], Nai Wei[1], Lei Xin[1,2] & Xiao-Dong Li[1,2] ✉

The predatory stink bug, *Eocanthecona furcellata*, is a crucial biological control agent used globally to manage agricultural pests. Here, we present the chromosome-level genome assembly for this economically important species to provide a foundational resource for understanding its unique predatory biology. By integrating 41.94 Gb of PacBio HiFi, 56.12 Gb of Illumina, and 68.62 Gb of Hi-C sequencing data, we constructed a 1.10 Gb assembly with excellent contiguity (contig N50 = 42.08 Mb; scaffold N50 = 144.30 Mb). A total of 98.85% of the assembled sequences were anchored into eight pseudochromosomes. Genomic analysis revealed that repetitive elements comprise 41.46% of the genome, and we annotated 16,880 protein-coding genes, 98.74% of which were functionally assigned. The quality of the assembly is confirmed by its high completeness (BUSCO score = 98.8%; Merqury QV = 46.33). This high-quality genomic resource will be invaluable for future research into the molecular mechanisms of predation, insecticide resistance, and for optimizing the application of *E. furcellata* in integrated pest management programs.

Background & Summary

The predatory stink bug, *Eocanthecona furcellata* (Wolff) (Hemiptera: Pentatomidae), is a polyphagous natural enemy widely distributed in tropical and subtropical regions. It serves as a significant biological control agent, preying on over 40 species of agricultural and forestry pests, particularly those from the orders Lepidoptera, Coleoptera, and Hymenoptera[1–3]. Its effectiveness in suppressing populations of major pests like *Spodoptera litura* and *Plutella xylostella* has established it as a valuable component in integrated pest management (IPM) programs[4,5]. The practical application of *E. furcellata* has spurred extensive research into optimizing its mass-rearing techniques, life table parameters, and reproductive success under various diets, aiming to enhance its field performance[1,4].

Given its crucial role in agricultural ecosystems, *E. furcellata* has become a focal point for toxicological and physiological studies. Research has demonstrated that its fitness, reproductive potential, and predatory efficacy can be significantly impacted by exposure to common insecticides such as λ-cyhalothrin and chlorpyrifos[6–8]. These findings highlight the delicate balance required to integrate chemical and biological control strategies and underscore the need to understand the underlying mechanisms of insecticide tolerance and detoxification. Moreover, studies on its venom composition, which includes components like serine proteases and their inhibitors, have begun to unveil the molecular basis of its potent predatory capabilities[9,10].

Despite extensive research into its ecology, behavior, and physiology, the genomic foundation underpinning these complex traits in *E. furcellata* remains largely unexplored. Previous molecular studies have primarily relied on transcriptomic data or focused on specific gene families, which, while valuable, cannot provide a complete picture of genome architecture, gene regulation, or the evolutionary adaptations related to predation

[1]Guangxi Key Laboratory of Sericulture Ecology and Applied Intelligent Technology, School of Chemistry and Bioengineering, Hechi University, Hechi, 546300, China. [2]Guangxi Key Laboratory of Sericulture Ecology and Applied Intelligent Technology, Guangxi Collaborative Innovation Center of Modern Sericulture and Silk, Guangxi Colleges Universities Key Laboratory of Exploitation and Utilization of Microbial and Botanical Resources, School of Chemistry and Bioengineering, Hechi University, Hechi, 546300, China. ✉e-mail: lxdong_627@163.com

and environmental resilience[10,11]. The absence of a high-quality, chromosome-level reference genome has been a significant bottleneck, limiting deeper investigations into the genetic basis of key biological processes such as venom evolution, insecticide resistance, and chemosensory perception[2,12].

The advancement of high-throughput sequencing technologies provides an unprecedented opportunity to address this knowledge gap. A chromosome-level genome assembly is an essential resource for elucidating the genetic underpinnings of an organism's unique biology, facilitating comparative genomics, and accelerating molecular breeding efforts for beneficial traits. For a key predatory insect like *E. furcellata*, a high-quality genome is pivotal for dissecting the genetic pathways involved in its predatory success and its response to environmental stressors, ultimately informing more sustainable and effective pest control strategies[3,8,13].

Here, we present the chromosome-level genome assembly for *E. furcellata*. The final genome size is 1.10 Gb, with 98.85% of the sequence anchored onto 8 chromosomes. The assembly exhibits high contiguity, with a contig N50 of 42.08 Mb and a scaffold N50 of 144.30 Mb. Repetitive elements constitute 41.46% of the genome. We successfully annotated 16,880 protein-coding genes. This high-quality genomic resource provides an essential foundation for exploring the genetic basis of predation, insecticide resistance, and developmental biology in *E. furcellata*. It will significantly advance our understanding of this economically important predator and offer new avenues for enhancing its application in biological control programs worldwide.

Methods

Sample collection and sequencing. To minimize genomic heterozygosity, a single adult male E. furcellata was selected from a continuously maintained laboratory colony originally established from individuals collected in Hainan, China. High-molecular-weight (HMW) genomic DNA was extracted from dissected thoracic muscle tissue using the DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany) following the manufacturer's protocol. To construct a comprehensive genomic dataset, we prepared three distinct sequencing libraries from this DNA sample. For long-read data, a SMRTbell library with an approximate 20 kb insert size was constructed and sequenced on a PacBio Sequel IIe platform to generate high-fidelity (HiFi) reads for primary contig assembly. For polishing and validation, a separate short-read library (350 bp insert size) was prepared for 150 bp paired-end sequencing on a BGISEQ-500 platform. Finally, to enable chromosome-level scaffolding, an *in situ* Hi-C library was prepared from thoracic muscle tissue. After chromatin cross-linking, digestion with the MboI restriction enzyme[14], and proximity ligation, the library was sequenced with 150 bp paired-end reads on the BGISEQ-500 platform.

To support high-quality gene annotation, total RNA was extracted from a pool of diverse tissues (head, thorax, abdomen, and legs) from another individual of the same population. Following poly(A) enrichment, a transcriptome library was prepared and sequenced on a BGISEQ-500 platform to generate 150 bp paired-end reads. All raw genomic and transcriptomic sequencing data underwent a rigorous quality control pipeline to remove adaptors and low-quality reads. The voucher specimen was deposited in the Insect Collection of Hechi University under accession ID Ef-2024-01.

Genome assembly. We constructed a high-quality reference genome for *Eocanthecona furcellata*. A hybrid sequencing strategy was employed, integrating three distinct data types. We generated 41.94 Gb of PacBio HiFi reads, providing approximately 38-fold coverage of highly accurate long-read data that formed the backbone of the assembly. These were supplemented by 56.12 Gb (52 × coverage) of Illumina short-read data for polishing and validation, and 68.62 Gb (64 × coverage) of *in situ* Hi-C data for chromosome-level scaffolding. The initial assembly was performed with hifiasm v0.19.8[15] with default parameters, leveraging the PacBio HiFi data to produce a set of highly contiguous and accurate contigs. This foundational assembly resulted in a total of 79 contigs with a contig N50 of 42.08 Mb, indicating exceptional continuity at the base level.

To achieve a chromosome-level assembly, the contigs were scaffolded using the comprehensive Hi-C dataset. The Hi-C reads were first rigorously filtered using HiC-Pro v3.1.025[16], and the valid interaction pairs were then mapped to the primary contigs with the Juicer pipeline v1.6[17] to generate a genome-wide contact matrix. This matrix revealed clear interaction patterns that were computationally resolved by HapHiC v1.0.5[18] using the 'cluster' and 'assembly' modules to order and orient the contigs into large-scale scaffolds, which were then manually curated using Juicebox v2.17[19]. The final Hi-C contact map demonstrated a strong diagonal signal and distinct, well-defined chromatin territories corresponding to individual chromosomes (Fig. 1), confirming the high accuracy and completeness of the scaffolding process. The final assembly yielded 16 scaffolds with a scaffold N50 of 144.30 Mb.

After binning into chromosomes, the final chromosomal scaffolds contain 63 sequence gaps between adjacent contigs. To characterize these unresolved regions, we mapped the original PacBio HiFi reads back to the assembly using quartet v1.2.5[20]. This analysis indicated that the total length of sequence required to close these gaps is estimated to be at least 157,150 bp. Furthermore, to evaluate the completeness of the chromosome ends, we searched for telomeric repeat sequences also using quartet v1.2.5[20]. Using the identified telomeric repeat motif (AATATC), our analysis predicted the presence of telomeres at both ends of two chromosomes (EoFu03 and EoFu04) and at a single end for five other chromosomes. In total, these predicted telomeric motifs were identified on seven of the eight pseudochromosomes, providing strong evidence for their high degree of structural completeness.

The resulting chromosome-level genome of *E. furcellata* spans a total of 1.10 Gb, with 1.08 Gb (98.85%) of the sequence successfully anchored and oriented into eight pseudochromosomes, consistent with the karyotypes of several other Pentatomidae species[21,22]. The overall GC content of the genome was determined to be 32.39%. When compared to other published genomes within the Pentatomidae family, the *E. furcellata* assembly demonstrates a substantial improvement in sequence continuity. While its genome size and scaffold N50 are comparable to those of species like *Aelia acuminata*[23] and *Piezodorus guildinii*[24], its contig N50 of 42.08 Mb is

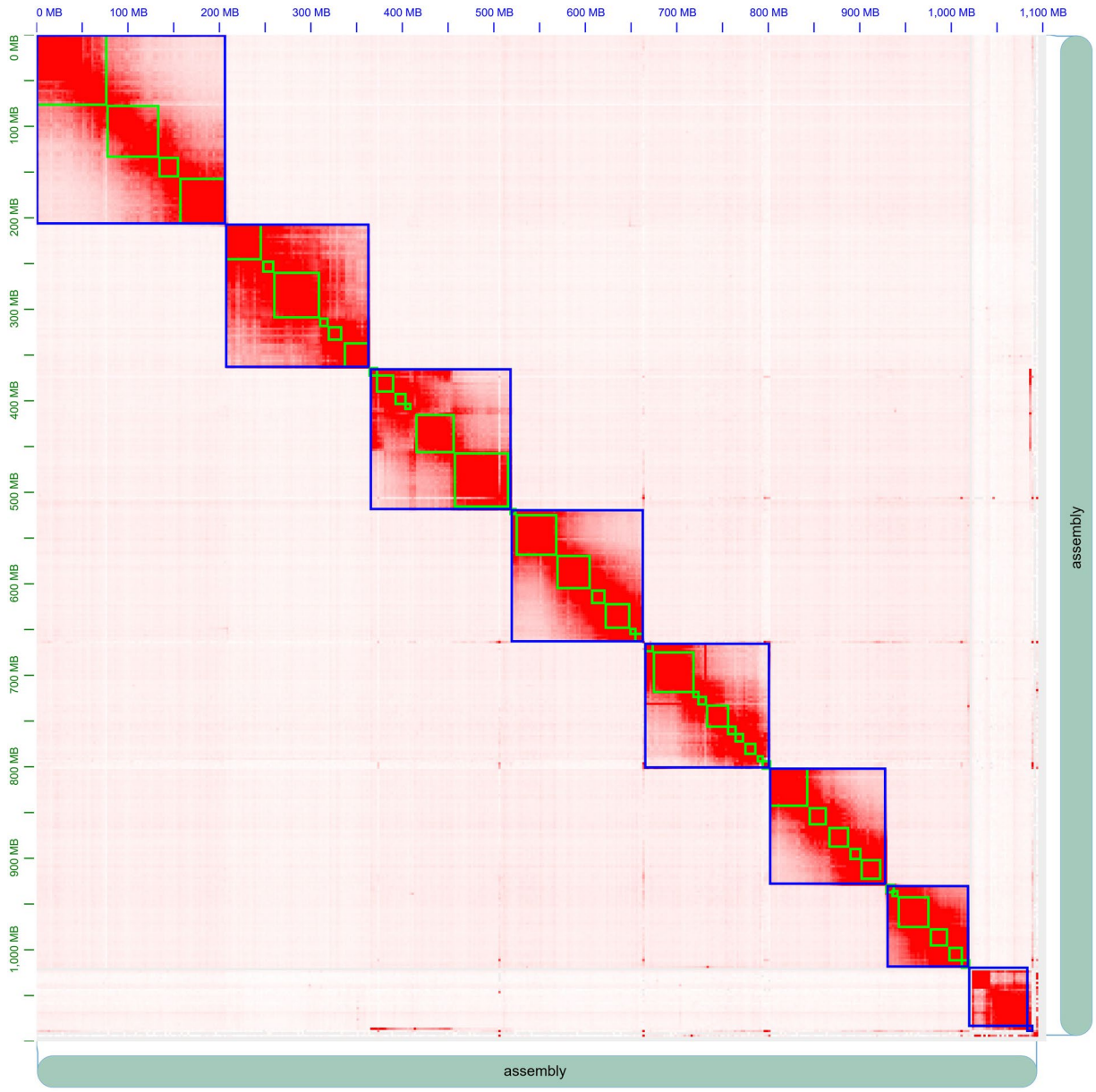

Fig. 1 Hi-C interaction heatmap validating the chromosome-level assembly of the *E. furcellata* genome. The plot visualizes the chromatin contact frequencies across the eight assembled pseudochromosomes. The strong signal intensity concentrated along the diagonal and the distinct, well-defined square territories confirm the structural integrity and the correct ordering and orientation of the scaffolds into final chromosomes.

significantly higher—over 5.7 times greater than the next best assembly—highlighting its exceptional quality as a valuable genomic resource (Table 1).

Genome annotation. The genomic landscape of *E. furcellata* was comprehensively characterized. Firstly, we conducted an in-depth analysis of its repetitive elements. Employing an integrated EarlGrey pipeline[25] that combined de novo identification using RepeatModeler v2.0.5[26] with homology-based searches against the Repbase database using RepeatMasker v4.1.5[27], our analysis revealed that a substantial portion of the genome consists of repetitive sequences. In total, these elements account for 41.46% of the assembly, occupying approximately 454.61 Mb (Fig. 2C, Table S2). The majority of these were classified as interspersed repeats (38.63%), with Long Interspersed Nuclear Elements (LINEs) being the most prevalent class, constituting 12.08% of the entire genome (Fig. 2E). This was followed by Long Terminal Repeats (LTRs) at 4.76% (Fig. 2D) and DNA transposons at 2.21% (Fig. 2F). Notably, a significant fraction of the repeats (19.84%) could not be classified into known families, suggesting the presence of novel or highly diverged repeat lineages specific to this predatory bug (Fig. 2G).

Following the masking of these repetitive regions, we performed a structural annotation to identify protein-coding genes. Our pipeline integrated three complementary streams of evidence: homology-based alignments using protein sets retrieved from hemipteran_odb12 database using miniport v0.18[28], transcriptome-guided

Assembly Metrics	*Aelia acuminata (GCA_911387785.2)*	*Rhaphigaster nebulosa (GCA_948150685.1)*	*Arma custos (GCA_048127545.1)*	*Piezodorus guildinii (GCA_023052935.1)*	*Eocanthecona furcellata*
Genome size (bp)	1,169,999,116	1,155,816,268	976,577,182	1,205,420,128	1,092,127,095
Chromosome sequnece No.	9	8	7	7	8
Genome sequence No.	79	75	24	807	65
GC content	33	32	33	32.5	32.39
Scaffold No.	79	75	24	807	18
Scaffold L50	3	3	3	4	4
Scaffold N50 (bp)	172,186,306	186,487,842	136,202,916	170,835,737	144,301,894
Scaffold N90 (bp)	93,278,539	133,601,672	110,891,642	118,461,799	94,292,339
Contig sequence No.	689	1,229	466	1,267	76
Contig L50	65	199	69	47	11
Contig N50 (bp)	4,690,255	1,716,941	4,490,611	7,358,154	42,082,235
No. Genes					16,680
Hifi Size/Depth					41.94 Gb (~38.1×)
DNA_NGS Size/Depth					56.12 Gb (~51.0×)
Hi-C Size/Depth					68.62 Gb (~62.4×)

Table 1. Summary of genome assembly parameters of *E. furcellata* and four published Pentatomidae genomes in the NCBI GenBank database.

predictions derived from mapped RNA-seq data using HISAT2 v2.2.1[29], and species-trained ab initio modeling with AUGUSTUS v3.5.0[30]. These evidence streams were synthesized using the GETA pipeline (https://github.com/chenlianfu/geta) to produce a high-quality, non-redundant gene set. This approach yielded a final count of 16,880 protein-coding genes, including 358 with alternative splicing isoforms (Fig. 2B). The structural characteristics of these genes revealed a mean length of 14.58 kb, with an average of 6.13 exons per gene. The coding sequences were relatively short (mean CDS length of 1.39 kb) compared to the expansive introns (mean intron length of 13.11 kb), indicating a genome architecture rich in non-coding regions.

To ascertain the functional potential of the predicted gene set, we conducted a comprehensive functional annotation using a dual-pronged strategy. First, all protein sequences were queried against the NCBI non-redundant (NR) protein database using DIAMOND blastp (v2.1.9) with sensitive mode[31]. Second, conserved domains and functional sites were identified using InterProScan v5.72-103.0[32] against the Pfam, CDD, SMART, and SUPERFAMILY databases, which also provided Gene Ontology (GO) term assignments. By merging the results from both analyses, we successfully assigned putative functions to 16,687 genes, representing 98.74% of the total annotated protein-coding gene set.

Data Records

The chromosome-level genome assembly of *E. furcellata* has been deposited in the National Center for Biotechnology Information (NCBI) GenBank database under the accession number GCA_052056905.1[33]. All raw sequencing data generated in this study have been made publicly available through the NCBI Sequence Read Archive (SRA) under the accession number SRP605685[34]. The PacBio HiFi long-read sequencing data used for primary contig assembly are accessible under accession number SRR34824520. Illumina short-read whole genome sequencing data used for error correction and heterozygosity assessment are available under accession number SRR34824518. The Hi-C sequencing data utilized for chromosome-level scaffolding can be accessed under accession number SRR34824519. Additionally, the RNA-sequencing data used for genome annotation and transcriptome analysis have been deposited under accession number SRR34824517. In addition, the final genome annotation files, including the coding sequences (CDS), protein sequences (PEP), gene feature format file (GFF3), and the comprehensive functional annotation table, have been deposited in Figshare and are accessible at https://doi.org/10.6084/m9.figshare.30157360[35].

Technical Validation

To validate the quality of the *E. furcellata* genome, we conducted a multi-faceted evaluation of its accuracy, completeness, and continuity. The structural integrity and fidelity of the assembly were first confirmed by mapping the original sequencing reads back to the final sequence using BWA v0.7.17[36] and minimap2 v2.30[37]. An overwhelming majority of the raw data aligned successfully, with a mapping rate of 99.47% for the Illumina short reads and 99.97% for the PacBio HiFi long reads, indicating that the assembly is a faithful and comprehensive representation of the genomic content. We further assessed the base-level accuracy using k-mer analysis with Merqury v1.3[38], which yielded an exceptionally high-quality value (QV) of 46.33, signifying an estimated error rate of less than one per 40,000 bases. The completeness of the gene space was then rigorously benchmarked using BUSCO v6.0.0[39] against the Hemiptera (hemiptera_odb12) single-copy ortholog set. This analysis revealed that 98.8% of the conserved genes were present and complete in our assembly, with 97.5% found as single-copy and 1.3% as duplicated. Only a minor fraction was identified as fragmented (0.1%) or were missing (1.1%). Collectively, the near-perfect read mapping rates, outstanding base-level accuracy, and high degree of gene completeness underscore the exceptional quality of our chromosome-level assembly, establishing it as a robust resource for subsequent genomic research.

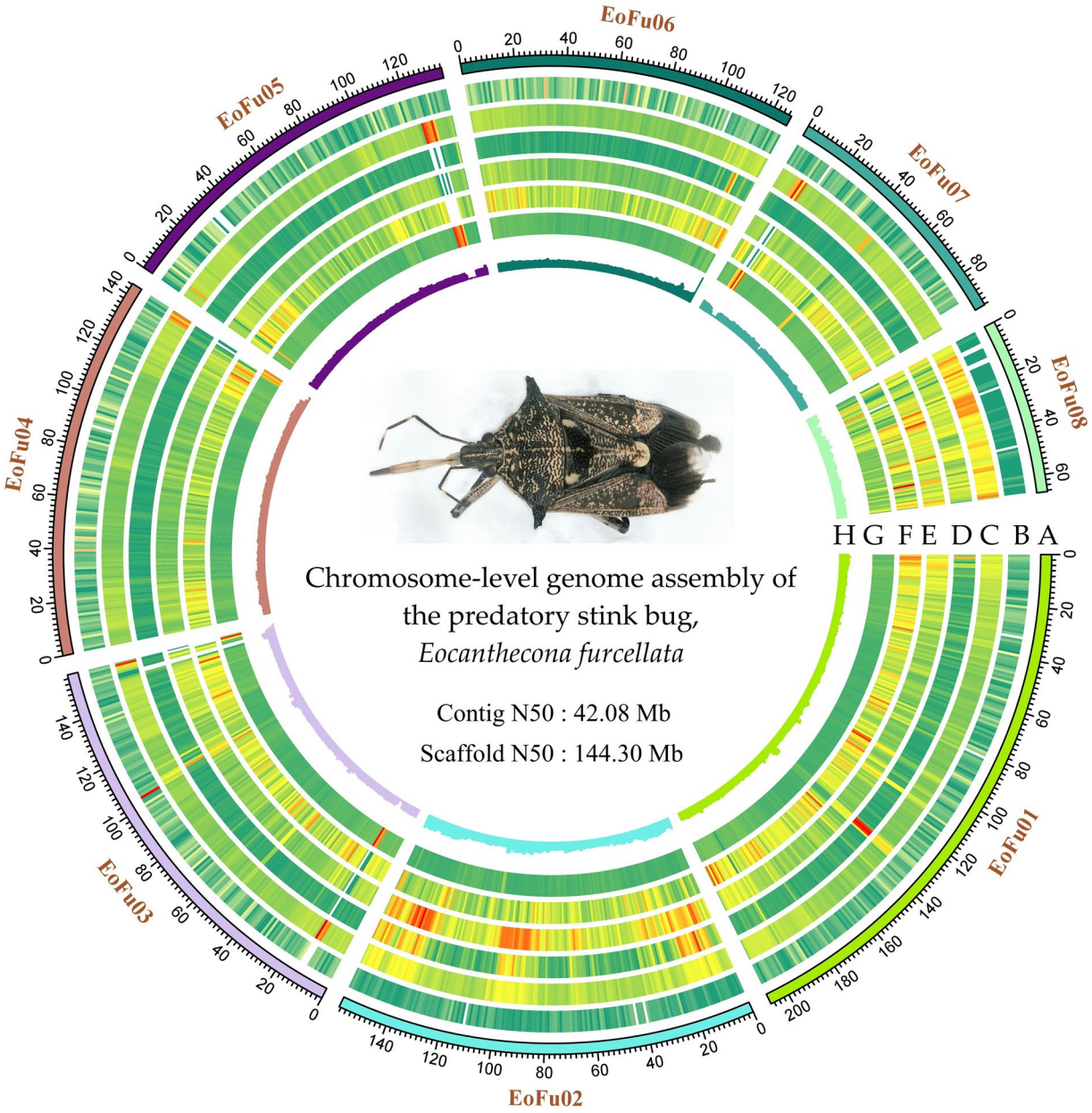

Fig. 2 Genomic landscape of the *E. furcellata* chromosome-level assembly. The circular plot displays key genomic features in concentric tracks. The tracks are ordered from the outermost circle inwards: (A) Chromosome Ideograms: The eight pseudochromosomes (EoFu01–EoFu08) are shown, with coordinate scales in megabases (Mb). (B) Gene Locations: A high-density tick plot where each individual mark indicates the precise position of an annotated protein-coding gene. (C) Total Repetitive Element Density: A heatmap illustrating the density of all identified repetitive elements. (D) LTR Retrotransposon Density: Heatmap showing the density of Long Terminal Repeat retrotransposons. (E) LINE Density: Heatmap showing the density of Long Interspersed Nuclear Elements. (F) DNA Transposon Density: Heatmap showing the density of DNA transposons. (G) Simple Repeat Density: Heatmap showing the density of simple repeats (microsatellites). (H) GC Content: A heatmap representing the percentage of Guanine and Cytosine. All density and content tracks (C–H) were calculated and plotted using a non-overlapping 100 kb window size. For all heatmap tracks, the colour gradient from green to yellow to red corresponds to low, medium, and high density or content values, respectively.

Data availability

The genome assembly and annotation data generated in this study have been deposited in the NCBI GenBank database under accession GCA_052056905.1. The raw sequencing reads are available in the NCBI Sequence Read Archive (SRA) under the accession number SRP605685[34]. The final genome annotation files are publicly available in the Figshare repository under https://doi.org/10.6084/m9.figshare.30157360[35].

Code availability

The entire analytical pipeline was implemented using well-established, open-source bioinformatics tools. This approach obviated the need for developing complex custom scripts and instead relied on validated, community-standard methods. Reproducibility was further enforced through systematic dependency management using the Conda environment system, which guarantees a consistent software stack across different computational infrastructures. For complete methodological clarity, the precise command-line instructions and all parameters employed for each data processing, assembly, and annotation step are explicitly provided in the Methods section, allowing for the independent verification and replication of our results.

Received: 4 August 2025; Accepted: 21 October 2025;
Published online: 05 December 2025

References

1. Yuan, J. *et al*. Demographic parameters and asynchronous ovary development in Eocanthecona furcellata (Hemiptera: Pentatomidae) reared on different diets. *Journal of economic entomology* **118**, 1051–1060, https://doi.org/10.1093/jee/toaf040 (2025).
2. Parveen, S., Ahmad, A., Brożek, J. & Ramamurthy, V. V. Morphological diversity of the labial sensilla of phytophagous and predatory Pentatomidae (Hemiptera: Heteroptera), with reference to their possible functions. *Zootaxa* **4039**, 359–372, https://doi.org/10.11646/zootaxa.4039.2.9 (2015).
3. Pan, Y. N. *et al*. Assessment of Suitable Reference Genes for qRT-PCR Normalization in Eocanthecona furcellata (Wolff). *Insects* **13**, https://doi.org/10.3390/insects13090773 (2022).
4. Tuan, S. J., Yeh, C. C., Atlihan, R. & Chi, H. Linking Life Table and Predation Rate for Biological Control: A Comparative Study of Eocanthecona furcellata (Hemiptera: Pentatomidae) Fed on Spodoptera litura (Lepidoptera: Noctuidae) and Plutella xylostella (Lepidoptera: Plutellidae). *Journal of economic entomology* **109**, 13–24, https://doi.org/10.1093/jee/tov265 (2016).
5. Pan, C. N. *et al*. Fitness implications of low-temperature storage for Eocanthecona furcellata (Hemiptera: Pentatomidae). *Journal of economic entomology* **117**, 1739–1752, https://doi.org/10.1093/jee/toae199 (2024).
6. Yao, Q. *et al*. Predatory stink bug, Eocanthecona furcellata (Wolff) responses to oral exposure route of λ-cyhalothrin via sex-specific modulation manner. *Pesticide biochemistry and physiology* **192**, 105381, https://doi.org/10.1016/j.pestbp.2023.105381 (2023).
7. Xu, S. *et al*. Sublethal effect of chlorpyrifos on predatory behavior and physiology of Eocanthecona furcellata (Hemiptera: Pentatomidae). *Journal of economic entomology* **117**, 156–166, https://doi.org/10.1093/jee/toad212 (2024).
8. Qiong, Y., Linfa, Q., Shu, X., Longyu, Y. & Bingxu, C. Detrimental Impact of λ-Cyhalothrin on the Biocontrol Efficacy of Eocanthecona furcellata by Affecting Global Transcriptome and Predatory Behavior. *Journal of agricultural and food chemistry* **70**, 1037–1046, https://doi.org/10.1021/acs.jafc.1c06237 (2022).
9. Zhang, M. *et al*. Functional analysis of serine protease EfSP1 in Eocanthecona furcellata Wolff (Hemiptera: Pentatomidae). *Pest management science* **81**, 4515–4529, https://doi.org/10.1002/ps.8809 (2025).
10. Zhang, M. *et al*. Identification and functional analysis of serine protease inhibitor gene family of Eocanthecona furcellata (Wolff). *Frontiers in physiology* **14**, 1248354, https://doi.org/10.3389/fphys.2023.1248354 (2023).
11. Pang, R. *et al*. Decreased cuticular penetration minimizes the impact of the pyrethroid insecticide λ-cyhalothrin on the insect predator Eocanthecona furcellata. *Ecotoxicology and environmental safety* **249**, 114369, https://doi.org/10.1016/j.ecoenv.2022.114369 (2023).
12. Zhao, H. *et al*. Observation of the fine structure of antennal sensilla of the stink bug, Eocanthecona furcellata (Hemiptera: Pentatomidae). *Micron (Oxford, England: 1993)* **150**, 103143, https://doi.org/10.1016/j.micron.2021.103143 (2021).
13. Yasui, H. Sequestration of host plant-derived compounds by geometrid moth, Milionia basalis, toxic to a predatory stink bug, Eocanthecona furcellata. *Journal of chemical ecology* **27**, 1345–1353, https://doi.org/10.1023/a:1010361125048 (2001).
14. Cardamone, F., Zhan, Y., Iovino, N. & Zenk, F. Chromosome Conformation Capture Followed by Genome-Wide Sequencing (Hi-C) in Drosophila Embryos. *Methods in molecular biology (Clifton, N.J.)* **2655**, 41–55, https://doi.org/10.1007/978-1-0716-3143-0_4 (2023).
15. Cheng, H. Y., Concepcion, G. T., Feng, X. W., Zhang, H. W. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170–+, https://doi.org/10.1038/s41592-020-01056-5 (2021).
16. Servant, N. *et al*. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biology* **16**, https://doi.org/10.1186/s13059-015-0831-x (2015).
17. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems* **3**, 95–98, https://doi.org/10.1016/j.cels.2016.07.002 (2016).
18. Zeng, X. F. *et al*. Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. *Nature Plants* **10**, https://doi.org/10.1038/s41477-024-01755-3 (2024).
19. Durand, N. C. *et al*. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Systems* **3**, 99–101, https://doi.org/10.1016/j.cels.2015.07.012 (2016).
20. Lin, Y. *et al*. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Horticulture research* **10**, uhad127, https://doi.org/10.1093/hr/uhad127 (2023).
21. Grozeva, S., Kuznetsova, V. G. & Anokhin, B. A. Karyotypes, male meiosis and comparative FISH mapping of 18S ribosomal DNA and telomeric (TTAGG) n repeat in eight species of true bugs (Hemiptera, Heteroptera). *Comparative cytogenetics* **5**, 355–374, https://doi.org/10.3897/CompCytogen.v5i4.2307 (2011).
22. Rebagliati, P. J. & Mola, L. M. Meiotic behavior and karyotypic variation in Acledra (Pentatomidae, Heteroptera). *Genetics and molecular research: GMR* **9**, 739–749, https://doi.org/10.4238/vol9-2gmr763 (2010).
23. Crowley, L. & Barclay, M. V. L. The genome sequence of the bishop's mitre shieldbug, Aelia acuminata (Linnaeus, 1758). *Wellcome open research* **6**, 320, https://doi.org/10.12688/wellcomeopenres.17400.1 (2021).
24. Walt, H. K., King, J. G., Towles, T. B., Ahn, S. J. & Hoffmann, F. G. Comparative Genomics and the Salivary Transcriptome of the Redbanded Stink Bug Shed Light on Its High Damage Potential to Soybean. *Genome biology and evolution* **16**, https://doi.org/10.1093/gbe/evae121 (2024).
25. Baril, T., Galbraith, J. & Hayward, A. Earl Grey: A Fully Automated User-Friendly Transposable Element Annotation and Analysis Pipeline. *Molecular Biology and Evolution* **41**, https://doi.org/10.1093/molbev/msae068 (2024).
26. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America* **117**, 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
27. Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics* Chapter 4, Unit 4.10, https://doi.org/10.1002/0471250953.bi0410s05 (2004).
28. Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics (Oxford, England)* **39**, https://doi.org/10.1093/bioinformatics/btad014 (2023).
29. Kim, D., Landmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* **12**, 357–U121, https://doi.org/10.1038/nmeth.3317 (2015).

30. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids research* **34**, W435–439, https://doi.org/10.1093/nar/gkl200 (2006).
31. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59–60, https://doi.org/10.1038/nmeth.3176 (2015).
32. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics (Oxford, England)* **30**, 1236–1240, https://doi.org/10.1093/bioinformatics/btu031 (2014).
33. Guan, D.-L. *Genbank* https://identifiers.org/ncbi/insdc.gca:GCA_052056905.1 (2025).
34. Guan, D.-L. *NCBI. Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP605685 (2025).
35. Guan *et al*. Genome annotation and protein-coding gene set for the chromosome-level assembly of the predatory stink bug, Eocanthecona furcellata. *Figshare* https://doi.org/10.6084/m9.figshare.30157360 (2025).
36. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **26**, 589–595, https://doi.org/10.1093/bioinformatics/btp698 (2010).
37. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)* **34**, 3094–3100, https://doi.org/10.1093/bioinformatics/bty191 (2018).
38. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* **21**, https://doi.org/10.1186/s13059-020-02134-9 (2020).
39. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol* **1962**, 227–245, https://doi.org/10.1007/978-1-4939-9173-0_14 (2019).

Acknowledgements

This research was funded by the Local Science and Technology Development Fund project guided by the central government (Heke ZY230301), the 2023 Annual Operation Subsidy Project of the Guangxi Key Laboratory of Sericulture Ecology and Applied Intelligent Technology (23-026-08), the Yizhou High-quality Development of Cocoon and Silk Industry Talent Introduction Project (HCZC2024-G3-810273-HZTG).

Author contributions

D.G. conceived the project and supervised the study. Y.Q. and S.Z. collected the samples. L.X. and S.Z. performed the experiments. Y.C.Q., Y.Q., and L.X. performed the computational analyses and data curation. N.W. performed formal investigation. N.W. and D.G. wrote the manuscript. L.X. and X.D.L. revised the manuscript. X.D.L and D.G. received fundings. All authors approved the final manuscript for submission.

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/s41597-025-06185-0.

Correspondence and requests for materials should be addressed to X.-D.L.