scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-scale genome assembly of *Helcystogramma triannulella* (Lepidoptera: Gelechiidae)

Ying Wang[1], Jinjun Cao[1]✉, Baoni Qin[1], Yannan Niu[1], Zhishan Cao[1], Dávid Murányi[2] & Weihai Li[1]✉

Helcystogramma triannulella **(Lepidoptera: Gelechiidae) is a highly destructive pest that primarily targets crops within the Convolvulaceae family, causing significant agricultural losses. Although this species is economically significant, there are still few genomic resources available for it, which limits our ability to understand its biology and develop effective control strategies. Here, we present the first chromosome-scale genome assembly for *H. triannulella* generated with PacBio HiFi and Hi-C sequencing. The results showed that the final genome assembly had a total length of 387.28 Mb, containing 3,068 contigs and 2,152 scaffolds, with contig N50 and scaffold N50 reaching 2.28 Mb and 12.75 Mb, respectively. Through Hi-C-assisted scaffolding, 91.77% of the assembly was assigned to 30 chromosomes. The BUSCO completeness value was 96.2%. Additionally, 49.02% of the genome sequences were identified as repetitive elements, and a total of 14,211 protein-coding genes were annotated. In summary, the *H. triannulella* genome serves as a valuable resource for advancing our understanding of its biology and facilitating the development of targeted pest management strategies.**

Background & Summary

The Gelechiidae family, one of the largest and most diverse groups of micro-moths within the Lepidoptera, comprises over 4,700 described species with a global distribution[1,2]. Members of this family are characterized by their small size and narrow wings, and they play significant ecological roles in various ecosystems[1]. However, many Gelechiidae species are also notorious agricultural pests, causing substantial economic damage to crops and stored food products. Notable examples include *Pectinophora gossypiella* (pink bollworm) and *Tuta absoluta* (tomato leafminer), both of which have caused significant agricultural losses worldwide[3].

Among these pests, *Helcystogramma triannulella* (Herrich-Schäffer, 1854) stands out as a highly destructive species affecting crops in the Convolvulaceae family (Fig. 1). This pest is widely distributed across Europe, Russia, the Caucasus, Central Asia, East Asia, and northern India[4,5]. The larvae of *H. triannulella* feed on the mesophyll tissue of host plants, leaving behind only the transparent leaf epidermis. This feeding behaviour severely impairs the plant's photosynthetic capacity, causing leaves to wither and, in severe cases, leading to host plant death[6]. In certain regions of China, crop damage rates attributed to *H. triannulella* have reached alarming levels of 60–85%[7], highlighting its significant agricultural impact. Despite its economic importance, current research on *H. triannulella* has primarily focused on its biological characteristics and control strategies[6,8–12], with limited exploration at the molecular level. To date, only ten chromosome-level genomes of Gelechiidae species are available in the NCBI databases (as of May 2025), and no genome assemblies have been reported for *H. triannulella* or any species within the *Helcystogramma* genus.

In this study, we integrated PacBio HiFi reads and Hi-C data to assemble the first chromosome-level genome of *H. triannulella*. This high-quality genome assembly provides a valuable resource for elucidating the molecular mechanisms underlying *H. triannulella*'s biology and offers a foundation for developing effective prevention and control strategies to mitigate its agricultural impact.

[1]Henan International Joint Laboratory of Taxonomy and Systematic Evolution of Insecta, Henan Institute of Science and Technology, Henan, 453003, China. [2]Department of Zoology, Eszterházy Károly Catholic University, Leányka u. 6, Eger, H-3300, Hungary. ✉e-mail: cjj1986108@163.com; lwh7969@163.com

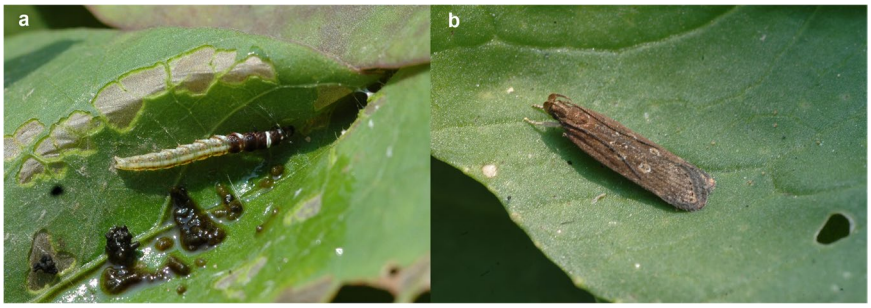

Fig. 1 Live larval (**a**) and adult (**b**) of *Helcystogramma triannulella*.

Library type	Platform	Library size (bp)	Data size (Gb)	Deep (X)	Application
short reads	DNBSEQ-T7	350	88.12	227.30	Genome survey
HiFi	PacBio SEQUELII	15,000	30.44	78.50	Genome assembly
Hi-C	DNBSEQ-T7	300-500	111.17	286.70	Chromosome construction
RNA-reads	DNBSEQ-T7	350	28.96	—	Annotation

Table 1. Summary of the sequence data for *Helcystogramma triannulella*.

Methods

Sample collection and sequencing. Specimens of *H. triannulella* were collected in September 2024 from Xinxiang City, China (35°16′48″N, 113°57′12″E) for whole genome sequencing. Field-collected *Ipomoea batatas* leaves exhibiting rolled leaves were transported to the laboratory. A portion of the leaves was used for larval collection, while the remainder was placed in wide-mouth bottles and maintained indoors until the emergence of *H. triannulella* adults. All collected samples were immediately flash-frozen in liquid nitrogen for storage at −80 °C. The larval samples were subsequently used for DNA sequencing to assemble the genome, while a pooled RNA sample including the different developmental stages (include larvae, pupae and adult) was used for transcriptome sequencing. In addition, the genomic DNA extracted from larval samples were also used for both Hi-C and PacBio HiFi sequencing.

The high-quality genomic DNA and total RNA were extracted using the CTAB method and Trizol reagent (Magen, Guangdong, China), respectively. The extracted DNA was evaluated for integrity and concentration using a NanoDrop 2000 spectrophotometer, Qubit 3.0 fluorometer, and 0.8% agarose gel electrophoresis to ensure suitability for downstream applications. Libraries were prepared for short-read sequencing using the VAHTS Universal Plus DNA Library Prep Kit for MGI (Vazyme, Nanjing, China), following the manufacturer's instructions. Hi-C libraries were prepared by digesting DNA with the MboI restriction enzyme and processed according to established protocols[13,14]. Both short-read and Hi-C libraries were sequenced on the BGISEQ DNBSEQ-T7 platform at Frasergen Bioinformatics (Wuhan, China). For long-read sequencing, PacBio HiFi libraries were first prepared using the SMRTbell Express Template Prep Kit 2.0. Subsequently, sequencing was performed on the PacBio Sequel II instrument with 8 M SMRT cells, acquiring one 30-hour movie per cell. This yielded a total of 30.44 Gb (78.50X coverage) of HiFi reads, 88.12 Gb (227.30X) of short reads, and 111.17 Gb (286.70X) of Hi-C reads (Table 1), providing a robust dataset for genome assembly. For transcriptome sequencing, RNA-seq libraries were prepared using the VAHTS Universal V10 RNA-seq Library Prep Kit (Vazyme, Nanjing, China) and sequenced on the BGISEQ DNBSEQ-T7 platform, generating 28.96 Gb of short-read RNA-seq data (Table 1). Subsequently, these transcriptomic data were utilized for genome annotation and the prediction of protein-coding genes, enabling a comprehensive understanding of the *H. triannulella* genome.

Genome survey and *de novo* assembly. To characterize the genomic features of *H. triannulella*, we first conducted a genome survey using short-read sequencing data to estimate genome size, repeat content, and heterozygosity. The frequency distribution of 17-mer sequences was analyzed using Jellyfish v. 2.3.1[15], and the results were further processed with GenomeScope v. 2.0[16] to derive genomic parameters. This analysis revealed an estimated genome size of 355.30 Mb, with a substantial proportion of repetitive sequences (39.4%) and a heterozygosity rate of 0.78% (Fig. 2).

For the *de novo* genome assembly, we utilized HiFi long-read sequencing data, which were processed using HiFiasm v. 0.16.1[17] with default parameters to generate a high-quality draft assembly. The sequence graphs in GFA format were converted to FASTA format using gfatools, ensuring compatibility with downstream analyses. The final *H. triannulella* genome assembly comprised 387.28 Mb of sequence, distributed across 3,068 contigs and 2,152 scaffolds, with the contig N50 and scaffold N50 reaching 2.28 Mb and 12.75 Mb, respectively (Table 2).

Chromosome assignment using Hi-C technology. To achieve a chromosomal-level genome assembly, we employed a comprehensive Hi-C-based scaffolding pipeline. Initially, Trimmomatic v. 0.40[18] was utilized to pre-process the Hi-C paired-end reads by removing low-quality bases and adapter sequences, ensuring high-quality input data for downstream analysis. The filtered reads were then aligned to the draft contigs using

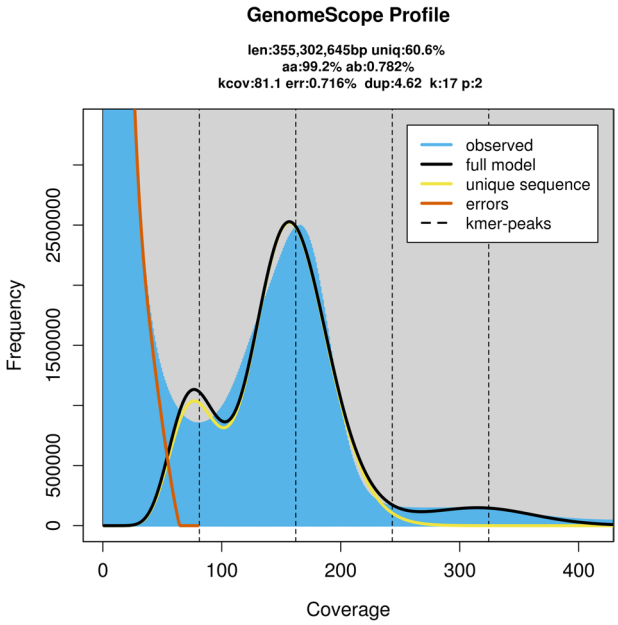

Fig. 2 K-mer (k = 17) distribution map of *Helcystogramma triannulella* genome.

Statistics	Value
Assembly length(bp)	387,277,015
Number of contigs	3,068
Contig N50 (bp)	2,284,609
Number of Scaffold	2,152
Scaffold N50 (bp)	12,745,528
anchor ratio (%)	91.77
BGI reads mapping rate (%)	96.09
BGI reads coverage (%)	98.27
HiFi reads mapping rate (%)	99.25
HiFi reads coverage (%)	99.28
BUSCO (%)	96.20
LTR assembly index	1.36
QV	27.14

Table 2. Statistics of Chromosome-level genome assembly of *Helcystogramma triannulella*.

Juicer (v3)[19] to calculate contact frequencies, which serve as the foundation for chromatin interaction mapping. Subsequently, 3d-DNA v. 180922[20] was applied with two iterative rounds of misjoin correction (-r2) using default parameters, refining the assembly by resolving potential structural errors. The oriented scaffolds generated from this step were further processed to construct interaction matrices using Juicer, which were visually inspected and manually corrected using Juicebox Assembly Tools v. 1.11.08[19]. In addition, the Hi-C-scaffolded genome sequence was ordered against the reference genome sequence (*Pectinophora gossypiella*, NCBI accession no. GCA_024362695.1) using Mummer v. 4.0.0rc1[21] alignment results. This iterative refinement process ensured the accuracy and continuity of the chromosomal scaffolds. The heatmap shows that on all 30 chromosomes, the interaction strength at diagonal positions is significantly higher than at non-diagonal regions, indicating good quality of chromosome assembly (Fig. 3A). The collinearity analysis indicated that the chromosomes of *H. triannulella* showed substantial collinearity with those of the pink bollworm *P. gossypiella* (Fig. 3C). Ultimately, we successfully generated the first chromosomal-scale high-quality assembly for *H. triannulella*, with the genome anchored onto 30 chromosomes (Fig. 3). The chromosomal lengths varied from 4.45 Mb to 25.60 Mb, collectively representing 91.77% of the total genome sequence (Table 3).

Assessment of assembly quality. The *H. triannulella* genome assembly was rigorously validated through multiple complementary assessment approaches. First, BUSCO (Benchmarking Universal Single-Copy Orthologue, v3.0.2)[22] analysis showed that 96.20% of the core conserved genes (5,085 out of 5,286 Lepidoptera_odb10) were complete in the assembly, demonstrating exceptional genome completeness (Table 4).

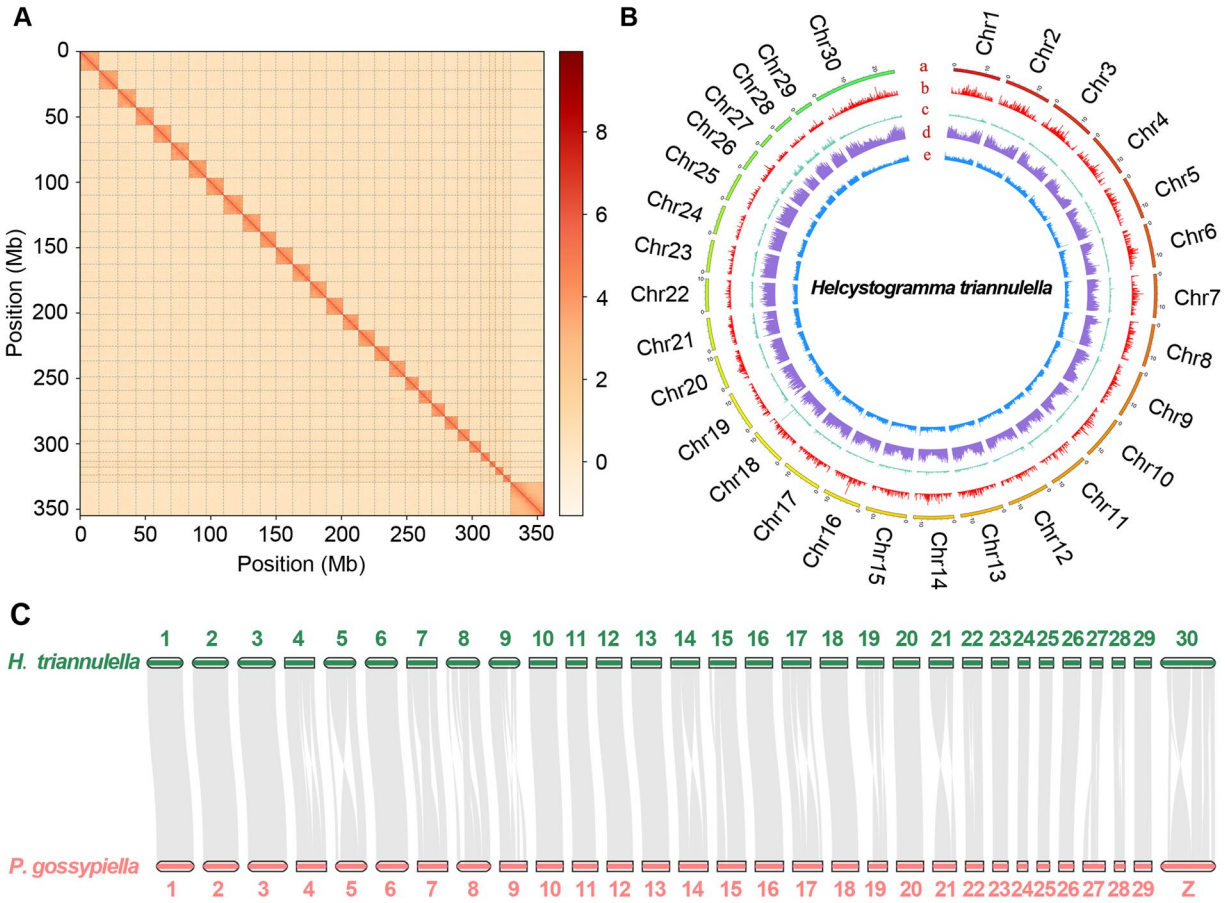

Fig. 3 Assembly of chromosome-level genome of the *Helcystogramma triannulella*. (**A**) The Hi-C interactive heatmap of *H. triannulella*. (**B**) The circle genome landscape of *H. triannulella*. a: 30 chromosomes assembly of *H. triannulella*; b: Gene frequency; c: Tandem repeat density; d: Transposable Element density; e: Guanine-cytosine (GC) content. (**C**) Collinearity relationship between *H. triannulella* and *Pectinophora gossypiella*.

Pseudomolecule	Contigs Length (bp)	Number of Contigs	Pseudomolecule	Contigs Length (bp)	Number of Contigs
Chr01	14,585,313	20	Chr16	11,824,234	21
Chr02	14,342,571	19	Chr17	12,414,123	24
Chr03	13,966,057	25	Chr18	11,246,059	18
Chr04	13,484,433	9	Chr19	12,225,589	50
Chr05	13,413,257	5	Chr20	10,234,572	31
Chr06	13,537,002	8	Chr21	9,963,502	50
Chr07	13,501,483	11	Chr22	10,018,866	32
Chr08	13,193,364	28	Chr23	9,899,931	26
Chr09	14,643,710	26	Chr24	9,018,721	13
Chr10	13,447,821	20	Chr25	8,741,992	33
Chr11	11,983,556	13	Chr26	6,681,889	31
Chr12	12,860,191	21	Chr27	4,452,997	71
Chr13	13,281,891	36	Chr28	5,998,648	67
Chr14	12,556,986	23	Chr29	5,549,802	66
Chr15	12,738,028	16	Chr30	25,604,516	27

Table 3. Statistics of Hi-C assembly results of *Helcystogramma triannulella*.

To further evaluate assembly accuracy, we assessed the reads re-mapping ratio and coverage using both BGI short reads and HiFi long reads. BGI short reads were aligned to the genome using BWA MEM v. 0.7.17[23] with default parameters, while HiFi long reads were mapped using Minimap2 v. 2.24[24] with the "-ax map-hifi" parameter. The resulting mapping rates of 96.09% and 99.25%, along with genome coverages of 98.27% and 99.28%, respectively, further confirmed the high quality and continuity of the assembly (Table 2).

Type	Number	Percent (%)
Complete BUSCOs (C)	5085	96.20
Complete and single-copy BUSCOs (S)	5045	95.44
Complete and duplicated BUSCOs (D)	40	0.76
Fragmented BUSCOs (F)	50	0.95
Missing BUSCOs (M)	151	2.86
Total BUSCO groups searched	5286	100

Table 4. BUSCO assessment results.

	Repbase		*de novo*		TE Proteins		Combined TEs	
Repeat type	Length (bp)	% in Genome	Length (bp)	% in Genome	Length (bp)	% in Genome	Length (bp)	% in Genome
DNA	14,014,957	3.61	76,664,687	19.77	593,584	0.15	83,404,701	21.51
LINE	13,973,169	3.60	64,966,127	16.76	6,840,173	1.76	73,930,530	19.07
SINE	18,063,731	4.66	28,611,296	7.38	0	0.00	43,992,974	11.35
LTR	2,860,965	0.74	29,561,762	7.62	2,650,024	0.68	30,791,978	7.94
Unknown	268,274	0.07	13,675,321	3.53	534	0.00	13,846,288	3.57
Total TE	48,353,901	12.47	168,990,493	43.58	10,079,148	2.60	190,063,932	49.02

Table 5. Summary statistics of repeat annotation in the *Helcystogramma triannulella* genome. Note: LINE, Long Interspersed Nuclear Elements; SINE, Short Interspersed Nuclear Elements; LTR, Long Terminal Repeats; Unknown, Repeat sequences that could not be classified by RepeatMasker; Total, Represents non-redundant data after removing overlaps between different classifications.

Given the inherent challenges of assembling highly repetitive genomic regions, we specifically evaluated the completeness of repetitive sequences using the Long Terminal Repeat (LTR) Assembly Index (LAI)[25]. The LAI value of 1.36 for the *H. triannulella* pseudomolecules indicates a reference-level assembly, underscoring the high completeness and accuracy of complex repetitive regions, including LTR retrotransposons. Additionally, we employed k-mer-based quality assessment (k = 19 bp) using the Merqury pipeline v. 1.3[26] with HiFi reads, which yielded a quality value (QV) score of 27.14 (Table 2). Collectively, these comprehensive evaluations demonstrate that the *H. triannulella* genome assembly achieves a high standard of quality, making it a reliable resource for downstream genomic and evolutionary analyses.

Repeat elements prediction. A comprehensive computational pipeline was used to identify and characterize repetitive sequences across the whole genome, including tandem repeats and transposable elements (TEs). Initially, Tandem Repeats Finder (TRF, v4.09.1)[27] with stringent parameters (2 7 7 80 10 50 2000) were used to annotate high-confidence tandem repeats. Subsequently, TE annotation was conducted through an integrated approach combining *de novo* prediction and homology-based identification at both nucleotide and protein levels. For DNA-level analysis, LTR retrotransposons (LTR-RTs) were initially identified using LTR_FINDER v. 1.0.7[28], followed by *de novo* repeat library construction through RepeatModeler v. 2.0.1[29], which generated a comprehensive repeat consensus database with detailed classification information. The resulting library, in conjunction with the curated Repbase TE database[30,31], was then subjected to RepeatMasker v. 4.1.2[32] for genome-wide TE identification. Concurrently, protein-level analysis was performed using RepeatProteinMask implemented in the RepeatMasker package, which employed WU-BLASTX against the TE protein database to enhance detection sensitivity. Our analysis revealed that repetitive elements constitute a substantial portion of the genome, with 190.06 Mb (49.02% of the genome assembly) identified as repetitive sequences. The TE landscape was dominated by DNA transposons, accounting for 21.51% of the total repetitive content. Other major TE classes included long interspersed nuclear elements (LINEs, 19.07%), short interspersed nuclear elements (SINEs, 11.35%), and long terminal repeats (LTRs, 7.94%), with a minor fraction (3.57%) remaining unclassified (Table 5).

Gene prediction and function annotation. We employed a comprehensive approach combining homologous, ab initio, and transcriptome-assisted annotation to predict the coding gene structures. For homologous annotation, we first performed Tblastn v. 2.11.0[33] comparisons between related species and the reference genome. The resulting aligned sequences and their corresponding proteins were subsequently filtered and processed using Exonerate v. 2.4.0[34] for precise alignment. For *de novo* annotation, we utilized both Augustus v. 3.4.0[35–37] and GlimmerHMM v. 3.0.4[38]. Regarding transcriptome analysis, we employed a dual strategy involving both genome-based and *de novo* assemblies. RNA-seq alignments were initially generated using HiSat2 v. .2.2.1[39], followed by transcript assembly through the genome-guided assembler Stringtie v. .2.1.7[40]. In parallel, we conducted *de novo* transcriptome assembly using Trinity v. 2.8.5[41]. All transcripts derived from RNA-seq were then integrated into a comprehensive transcriptome database following the PASA pipeline v. 2.4.1[42]. To consolidate our findings, we used Maker v. 3.01.03[43] to integrate the predicted gene sets into a non-redundant, comprehensive, and reliable gene collection. Finally, we refined these predictions through the PASA pipeline v. 2.4.1[42],

Annotation Type	Number	Percent (%)
InterPro	10,570	74.38
GO	7,240	50.95
KEGG	13,767	96.88
SwissProt	9,381	66.01
TrEMBL	13,706	96.45
NR	13,988	98.43
Total Annotated genes	13,989	98.44
Predicted protein-coding genes	14,211	—

Table 6. Number of predicted genes of *Helcystogramma triannulella* functionally annotated by using indicated databases.

Type	Counts	Average Length (bp)	Total Length (bp)	Percentage of sequence (%)
miRNA	74	82.96	6,139	0.0016
tRNA	17,993	72.69	1,307,966	0.3377
rRNA	52	431.54	22,440	0.0058
18S	3	1,872.00	5,616	0.0015
28S	3	3,852.67	11,558	0.0030
5S	46	114.48	5,266	0.0013
snRNA	65	149.22	9,699	0.0025
CD-box	8	175.88	1,407	0.0004
HACA-box	1	138.00	138	0.0000
splicing	56	145.61	8,154	0.0021

Table 7. Summary statistics of Non-coding RNA annotation results.

which enabled the addition of untranslated region (UTR) annotations and alternative splicing isoform models. Finally, we predicted a total of 14,211 protein-coding genes in the genome (Table 6). To functionally annotate the predicted protein-coding genes in the *H. triannulella* genome, we employed a comprehensive approach utilizing multiple well-established databases. Gene functions were inferred according to the best match of the alignments to the NCBI Non-Redundant (NR), KEGG database[44], Gene Ontology (GO)[45], TrEMBL[46] and Swiss-Prot[46] protein databases using Diamond BLASTP v2.0.7[47] with an E-value threshold of 1E-5. Additionally, protein domains were annotated using InterProScan v. 5.50-84.0[48] based on the InterPro[49] protein database, which integrates predictive models from multiple resources to provide a robust annotation framework. This multi-database approach enabled the functional annotation of 13,989 genes, representing 98.44% of the predicted protein-coding genes in the *H. triannulella* genome. Specifically, 74.38% of the genes were annotated in InterPro, 50.95% in GO, 96.88% in KEGG, 66.01% in SwissProt, 96.45% in TrEMBL, and 98.43% in NR (Table 6).

Annotation of non-coding RNA genes. The tRNAscan-SE (v2.0.9) algorithms[50] and RNAmmer v. 1.2[51] was used to predict tRNA genes and rRNA sequences, respectively. MiRNAs and snRNAs were identified by Infernal v. 1.1.2[52] software against the Rfam v. 14.6 database[53] with default parameters. This integrated approach resulted in the identification of 74 miRNAs, 17,993 tRNAs, 52 rRNAs, and 65 snRNAs in the *H. triannulella* genome. Notably, tRNAs were the most abundant ncRNA class, representing 0.3377% of the total genome length (Table 7).

Data Records

The genomic data of *H. triannulella* have been deposited in NCBI under the following accession numbers: PRJNA1215029[54] (BioProject) and SAMN46392533[55] (BioSample). Raw sequencing data, including BGI, PacBio, Hi-C, and transcriptome reads, are available in the NCBI Sequence Read Archive with accession numbers SRR33279372-SRR33279375[56–59].

BGI, PacBio, Hi-C, and transcriptome raw data have been deposited in the NCBI Sequence Read Archive with accession numbers SRR33279372-SRR33279375[56–59]. The genome assembly of *H. triannulella* has been submitted to NCBI GenBank with the accession number JBNFXN000000000[60]. Additionally, the genome assembly and annotation results are publicly accessible via the Figshare database[61].

Technical Validation

Two methods were employed to evaluate genome assembly quality. Initially, assembly completeness was assessed with BUSCO v3.0.2 against a reference set of Lepidoptera genes, which comprises 5,286 entries. This analysis revealed that 96.2% of the complete BUSCOs were successfully incorporated into the assembled genome, as detailed in Table 4. Subsequently, the accuracy of the assembly was gauged by calculating the mapping rates. The observed mapping rates were notably high, with PacBio reads at 99.25% and BGI reads at 96.09%, respectively. The results of these evaluations together demonstrate the high quality of the genome assembly generated in this study.

Code availability

All bioinformatic analyses were performed following the guidelines of the relevant software. Detailed parameters are provided in the Methods section, and no specific codes were used in this study.

Received: 15 May 2025; Accepted: 18 August 2025;
Published online: 01 September 2025

References

1. Karsholt, O., Mutanen, M., Lee, S. & Kaila, L. A molecular analysis of the Gelechiidae (Lepidoptera, Gelechioidea) with an interpretative grouping of its taxa. *Syst. Entomol.* **38**, 334–348 (2013).
2. Huemer, P. *et al*. DNA barcode library for European Gelechiidae (Lepidoptera) suggests greatly underestimated species diversity. *Zookeys* **921**, 141–157 (2020).
3. Zhang, B. C. *Index of Economically Important Lepidoptera*. CAB International, Wallingford. (1994).
4. Ponomarenko, M. G. Catalogue of the subfamily Dichomeridinae (Lepidoptera, Gelechiidae) of Asia. *Far. East. Entomol* **50**, 1–67 (1997).
5. Ma, L., Dong, W. W., Jiang, G. F. & Wang, X. The complete mitochondrial genome of *Brachmia macroscopa* (Lepidoptera: Gelechiidae) and its related phylogenetic analysis. *J. Insect Sci.* **16**, 9 (2016).
6. Ma, L., Wang, X., Liu, Y., Su, M.-Z. & Huang, G.-H. Temperature effects on development and fecundity of *Brachmia macroscopa* (Lepidoptera: Gelechiidae). *PLoS One.* **12**, e0173065 (2017).
7. Wen, L. & Wang, C. C. Biological features, occurrence regulation and control technology of *Brachmia macroscopa* Meyrick (Lepidoptera: Gelechiidae). *China Veget.* **13**, 28–29 (2010).
8. Hirano, C. & Muramoto, H. Effect of age on mating activity of the sweet potato leaf folder, *Brachmia macroscopa* (Lepidoptera: Gelechiidae). *Appl. Entomol. Zool.* **11**, 154–159 (1976).
9. Ma, L. *et al*. Developmental and reproductive response of *Brachmia macroscopa* (Lepidoptera: Gelechiidae) to three host plants. *Sci. Rep.* **8**, 8999 (2018).
10. Wang, J. L. & Tan, R. R. Effect of temperature developmental duration and feeding amount in *Brachmia macroscopa* Meyrick. *Changjiang Veget.* **4**, 75–77 (2011).
11. Khurana, A. D., Verma, A. N. & Bhanot, J. P. Chemical control of leaf feeding caterpillar *Brachmia macroscopa* Meyrick on sweet potato. *Haryana Agricultural University Journal of Research.* **4**, 173–175 (1974).
12. Muramoto, H., Souda, E. & Hirano, C. Laboratory bioassay of the sex pheromone of the female sweet potato leaf folder, *Brachmia macroscopa* (Lepidoptera: Gelechiidae). *Jpn. J. Appl. Entomol. Z.* **20**, 77–80 (1976).
13. Rao, S. S. *et al*. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell.* **159**, 1665–1680 (2014).
14. Xie, T. *et al*. De novo plant genome assembly based on chromatin interactions: A case study of Arabidopsis thaliana. *Mol. Plant.* **8**, 489–492 (2015).
15. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics.* **27**, 764–770 (2011).
16. Vurture, G. W. *et al*. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
17. Cheng, H. *et al*. Haplotype-resolved assembly of diploid genomes without parental data. *Nat. Biotechnol.* **40**, 1332–1335 (2022).
18. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* **30**, 2114–2120 (2014).
19. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
20. Dudchenko, O. *et al*. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
21. Marçais, G. *et al*. MUMmer4: A fast and versatile genome alignment system. *PLoS Computat Biol* **14**, e1005944 (2018).
22. Simão, F. A. *et al*. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics.* **31**, 3210–3212 (2015).
23. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at http://arxiv.org/abs/1303.3997v2 (2013).
24. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* **34**, 3094–3100 (2018).
25. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
26. Rhie, A. *et al*. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
27. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
28. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
29. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci.* **117**, 9451–9457 (2020).
30. Jurka, J. Repbase Update - a database and an electronic journal of repetitive elements. *Trends Genet.* **16**, 418–420 (2000).
31. Jurka, J. *et al*. Repbase update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
32. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics.* **4**, 10 (2009).
33. Gertz, E. M. *et al*. Composition-based statistics and translated nucleotide searches: Improving the TBLASTN module of BLAST. *BMC Biol.* **4**, 41 (2006).
34. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics.* **6**, 31 (2005).
35. Stanke, M. *et al*. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, W309–W312 (2004).
36. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465–W467 (2005).
37. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
38. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open-source ab initio eukaryotic gene-finders. *Bioinformatics.* **20**, 2878–2879 (2004).
39. Kim, D. *et al*. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
40. Kovaka, S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).
41. Grabherr, M. G. *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
42. Haas, B. J. *et al*. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **21**, 5654–5666 (2003).

43. Cantarel, B. L. *et al.* MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).
44. Kanehisa, M. *et al.* KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109–D114 (2012).
45. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
46. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).
47. Buchfink, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods.* **18**, 366–368 (2021).
48. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
49. Mitchell, A. *et al.* The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.* **43**, D213–D221 (2015).
50. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
51. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108 (2007).
52. Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics.* **25**, 1335–1337 (2009).
53. Kalvari, I. *et al.* Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.* **49**, D192–D200 (2021).
54. *NCBI BioProject* https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1215029 (2025).
55. *NCBI BioSample* https://www.ncbi.nlm.nih.gov/biosample/SAMN46392533 (2025).
56. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33279372 (2025).
57. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33279373 (2025).
58. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33279374 (2025).
59. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33279375 (2025).
60. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBNFXN000000000 (2025).
61. Chromosome-level genome assembly of *Helcystogramma triannulella* (Lepidoptera: Gelechiidae). *Figshare* https://doi.org/10.6084/m9.figshare.28953890 (2025).

Acknowledgements

This research was funded by the Natural Science Foundation of Henan Province (252300420178).

Author contributions

Y.W., J.C. and W.L. contributed to the research design. Y.W., Y.N. and B.Q. collected the samples. Y.W., J.C. and B.Q. analyzed the data. Y.W., J.C. and W.L. performed data analysis and wrote the manuscript. Z.C. and D.M. improved the manuscript. All authors have 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 J.C. or W.L.