scientific data

OPEN

DATA DESCRIPTOR

A chromosome-scale genome assembly of the leaf roller, *Eumorphobotys obscuralis* (Lepidoptera: Pyralidae)

Shu-Zhen Wang[1,2], Jun-Ran Zhao[1], Yu-Mei Dong[1], Wei Zhang[3], Yue Ying[3], Jin-Ping Shu[3] ✉, Jin-Ming Zhang[2] ✉ & Peng-Jun Zhang[1] ✉

The leaf roller, *Eumorphobotys obscuralis*, is a highly destructive pest that causes severe defoliation, ultimately reducing plant vigor, leading to culm death, and resulting in significant economic and ecological losses. Despite its substantial impact, effective prevention and control strategies remain lacking, further exacerbating the damage caused by this species. To address the critical gap in genomic resources, we generated a high-quality chromosome-scale genome assembly for *E. obscuralis* using a combination of Illumina short reads, PacBio HiFi long reads, and Hi-C chromatin interaction data. The final assembly spans 701.9 Mb across 31 chromosomes, including the single sex chromosome Z. Key assembly metrics include a contig N50 of 1.73 Mb and a scaffold N50 of 23.70 Mb, with an estimated genome completeness of 96.7%. Genome annotation identified 15,855 protein-coding genes and revealed that repetitive sequences account for 58.9% of the genome. This genomic resource lays a solid foundation for exploring the genetic mechanisms underlying the pest's biology and adaptation, and serves as a valuable reference for the development of targeted pest management strategies.

Background & Summary

The leaf roller, *Eumorphobotys obscuralis* (Lepidoptera: Pyralidae), first described by Aristide Caradja in 1925, is a member of the family Crambidae and is native to China[1]. It poses a significant threat to bamboo, particularly young plants, as its larvae feed on the leaves. Outbreaks of *E. obscuralis* have been frequently reported in China, India, Japan, and Korea, where they cause severe defoliation, leading to reduced plant vigor and even the death of bamboo culms.

Bamboo plays a vital role in the daily lives of millions across tropical regions, offering substantial environmental, social, and economic benefits[2,3]. However, damage caused by *E. obscuralis* has emerged as one of the primary threats to bamboo cultivation. During outbreaks, larvae feed aggressively on bamboo foliage, leading to the potential death of entire bamboo stands and reducing biomass by 35–50%. Such damage not only impairs bamboo shoot and whip growth in the following season but can also have lasting effects over several years. Currently, no effective prevention or control strategies exist, allowing damage from *E. obscuralis* to intensify year after year[4].

The absence of a high-quality genome assembly for *E. obscuralis* has hindered in-depth studies of its biology and management. To address this gap, we generated a chromosome-scale genome assembly using a combination of Illumina short reads, PacBio HiFi long reads, and high-throughput chromatin conformation capture (Hi-C) data. This genomic resource provides a critical foundation for advancing research on *E. obscuralis* by enabling investigations into its genetic architecture and supporting the development of targeted pest control strategies. With this assembly, researchers can now explore previously inaccessible areas of the genome, paving the way for novel insights and applications.

[1]College of Life and Environmental Sciences, Hangzhou Normal University, Hangzhou, 311121, China. [2]Institute of Plant Protection and Microbiology, Zhejiang Academy of Agricultural Sciences, Hangzhou, 310018, China. [3]Research Institute of Subtropical Forestry, Chinese Academy of Forestry, Hangzhou, 310018, China. ✉e-mail: jpshu@caf.ac.cn; zhanginsect@163.com; peng_junzhang@hotmail.com

Library type	Platform	Data size (Gb)	Depth[a] (X)	Average length (bp)
WGS short reads	Illumina HiSeq X-Ten	41.44	~63.7	150
WGS long reads	Pacbio Sequel II	21.10	~32.5	16,532
Hi-C	Illumina HiSeq X-Ten	55.80	~85.8	150
RNA-Seq	Illumina HiSeq X-Ten	11.35	—	150

Table 1. Summary statistics of sequencing data generated for *E. obscuralis* genome assembly and annotation. [a]For the convenience of calculation, the genome size of the *E. obscuralis* is set to 650 Mb.

Methods

Sample preparation. The original population of *E. obscuralis* was collected from Anji (30°41′N, 119°30′E) in Zhejiang Province, China. After collection, the population was maintained in the laboratory for approximately three months (roughly two generations) under controlled environmental conditions: 26 ± 2 °C, 60 ± 5% relative humidity, and a 14:10 hour light/dark photoperiod. During this period, both moths and larvae were reared on bamboo leaves.

Genomic DNA and RNA sequencing. Due to the small body size of *E. obscuralis*, a single adult did not yield sufficient DNA for high-throughput library construction. Therefore, female adults from the same laboratory-reared population were selected and pooled for genome sequencing. All individuals used for Illumina short-read, PacBio HiFi, and Hi-C sequencing originated from this cultured colony, which was established from a single field collection. Although the selected individuals were not clonal or derived from a single female, they belonged to the same generation and were maintained under identical rearing conditions to minimize variation.

Genomic DNA was extracted using the cetyltrimethylammonium bromide (CTAB) method. The quality of the extracted DNA was evaluated using three approaches: (1) agarose gel electrophoresis to assess DNA integrity and detect potential RNA contamination; (2) a NanoDrop ND-2000 spectrophotometer (Thermo Scientific, Wilmington, USA) to evaluate DNA purity; and (3) a Qubit® 4.0 Fluorometer (Life Technologies, CA, USA) to accurately quantify DNA concentration. Only samples meeting the following quality criteria were considered suitable for downstream applications: an OD260/280 ratio between 1.8 and 2.0, a concentration ≥ 50 ng/μL, and the presence of high-molecular-weight DNA with minimal RNA contamination.

For Illumina sequencing, genomic DNA (gDNA) was extracted from approximately 10 female *E. obscuralis* adults. Library preparation was performed using the TruSeq Nano DNA HT Sample Preparation Kit (Illumina, USA) following the manufacturer's protocol, with sample-specific indices incorporated during the process. Briefly, genomic DNA was fragmented to an average size of ~350 bp using ultrasonication. The resulting fragments underwent end-repair, A-tailing, adapter ligation, and PCR amplification, followed by purification. The quality of the constructed libraries was assessed in two steps: DNA concentration was initially measured using a Qubit® 4.0 Fluorometer (Life Technologies, CA, USA), and the insert size distribution was evaluated using an Agilent 2100 Bioanalyzer. Libraries were considered acceptable if the insert size fell within the expected range of ~300–500 bp. Quantitative PCR (qPCR) was then used to accurately determine the effective concentration of each library, with a threshold of >2 nM required for sequencing. Qualified libraries were sequenced on the Illumina NovaSeq 6000 platform (San Diego, CA, USA) using 150 bp paired-end reads, conducted by Smartgenomics Technology Institute (Tianjin, China). A total of 41.44 Gb of high-quality Illumina short-read data was generated (Table 1).

For long-read sequencing, genomic DNA (gDNA) was extracted from approximately 20 female *E. obscuralis* adults. DNA quality and concentration were assessed using a NanoDrop spectrophotometer and a Qubit 4.0 Fluorometer. High-quality DNA was then purified using AMPure PB beads (PacBio 100-265-900) and used for library construction. A PacBio SMRTbell (single-molecule real-time) library was prepared using the SMRTbell® Express Template Prep Kit 3.0 (Pacific Biosciences, USA). Briefly, gDNA was sheared to an average size of ~15 kb using the Megaruptor instrument (Diagenode, B06010001), followed by enrichment and purification of target-size fragments using magnetic beads. The sheared DNA underwent damage repair and end-repair, and sequencing adapters were ligated to both ends. Size selection of target fragments was performed using the BluePippin system (Sage Science, USA). The size distribution and concentration of the final library were assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies, USA). Libraries that passed quality control were loaded onto SMRT Cells and sequenced on the Sequel II/IIe platform (Pacific Biosciences, CA, USA) with a 30-hour runtime, generating 21.1 Gb of clean long-read data (Table 1).

For Hi-C sequencing, approximately 20 female *E. obscuralis* adults were fixed in 1% formaldehyde to cross-link DNA and associated proteins, thereby preserving native chromatin interactions. The cross-linking reaction was quenched with 0.125 M glycine, and tissues were homogenized to isolate intact nuclei. Chromatin was then digested with the restriction enzyme HindIII, producing cohesive DNA ends. These ends were repaired and labeled with biotin-14-dATP using DNA polymerase. Next, T4 DNA ligase was used to ligate spatially adjacent DNA fragments, capturing three-dimensional genomic interactions. Following proximity ligation, Proteinase K treatment was applied to reverse cross-links and degrade proteins. The DNA was purified via phenol-chloroform extraction and ethanol precipitation. Purified DNA was then sheared to an average fragment size of 300–500 bp using ultrasonication. Biotin-labeled fragments were selectively enriched using Dynabeads® M-280 Streptavidin (Life Technologies, USA). The Hi-C library was completed by ligating Illumina sequencing adapters, followed by PCR amplification. Sequencing was performed on the Illumina NovaSeq 6000 platform using a 150 bp paired-end (PE150) strategy. A total of 55.8 Gb of raw Hi-C data was generated (Table 1).

Chromosome	Contig number	Sequence length (bp)
Chr1	20	10,667,155
Chr2	16	23,842,754
Chr3	17	23,692,200
Chr4	17	18,672,942
Chr5	13	20,726,136
Chr6	22	17,763,802
Chr7	14	26,141,265
Chr8	22	25,686,113
Chr9	21	27,581,430
Chr10	13	23,702,690
Chr11	16	18,244,496
Chr12	13	21,985,269
Chr13	26	18,664,030
Chr14	26	31,059,042
Chr15	22	27,159,171
Chr16	13	16,269,721
Chr17	22	30,352,260
Chr18	13	10,602,170
Chr19	23	24,095,560
Chr20	26	24,092,830
Chr21	21	22,813,752
Chr22	21	18,461,828
Chr23	25	29,037,521
Chr24	30	21,170,217
Chr25	20	22,451,915
Chr26	24	22,312,397
Chr27	24	26,176,900
Chr28	14	23,786,931
Chr29	23	25,441,132
Chr30	36	14,361,459
Chr31	30	13,742,623

Table 2. Summary of contig counts and total lengths for each assembled chromosome in the *E. obscuralis* genome.

For transcriptome sequencing, total RNA was extracted from a single adult female *E. obscuralis* using the Qiagen RNeasy Plus Mini Kit, following the manufacturer's protocol. RNA integrity was assessed using an Agilent 2100 Bioanalyzer, and only samples with an RNA Integrity Number (RIN) greater than 7.0 were used for sequencing. mRNA libraries were prepared using the Illumina TruSeq Stranded mRNA Library Prep Kit and sequenced on the Illumina NovaSeq 6000 platform with 150 bp paired-end reads (2×150 bp). Sequencing was performed by Smartgenomics Technology Institute (Tianjin, China), generating a total of 11.35 Gb of high-quality transcriptomic data (Table 1).

Estimation of genomic parameters. Illumina raw reads were quality-filtered using fastp[5] (v0.23.2) to remove: (1) reads containing adapter contamination (>5 bp of adapter sequence); (2) reads with low sequencing quality (≥15% of bases with Q-scores < 19); (3) reads with a high proportion of ambiguous bases (N content > 5%); and (4) reads whose paired-end counterpart had been discarded in earlier filtering steps. Genome characteristics were assessed using a k-mer-based approach. K-mer frequency distribution, which reflects genome complexity and typically follows a Poisson distribution, was analyzed using Jellyfish[6] (v1.0.0) with a k-mer size of 17 (-m 17). Based on this distribution, genome size and heterozygosity were estimated using GCE[7] (v1.0.2). The analysis revealed a total of 36,634,361,699 k-mers, a peak k-mer depth of 55, and an estimated genome size of 666.08 Mb. After error correction, the modified genome size was refined to 653.72 Mb. The heterozygosity rate was estimated at 1.87%, and the duplication rate was 49.17%.

Genome assembly. The PacBio SMRT Analysis software package (https://www.pacb.com) was used to perform quality control on continuous long reads generated from SMRT sequencing. Filtering criteria included: (1) reads shorter than 50 bp; (2) reads with a read quality (RQ) score below 0.8; (3) reads containing self-ligated adapters; and (4) adapter trimming. High-fidelity circular consensus sequencing (CCS) reads were then generated using SMRT Link v9.0 with parameters--min-passes = 3 and--min-rq = 0.99. The CCS reads were assembled using Hifiasm[8] (https://github.com/chhylp123/hifiasm), producing a primary assembly of contigs. To assess the utility of Hi-C data, raw Hi-C reads were aligned to the contigs using HiCUP[9], allowing for the evaluation of valid

Features	*E. obscuralis*	*A. transitella*	*P. interpunctella*
Assembly accession	GCA_048418795.1	GCF_032362555.1	GCF_027563975.2
Genome size (Mb)	701.95	327.4	291.3
Karyotype	30 + ZW	30 + ZW	30 + ZW
Number of contigs	685	32	71
Number of scaffolds	73	32	45
Contig N50 (kb)	1,725.37	11,369.41	7,837.05
Scaffold N50 (Mb)	23.70	11.37	10.31
Number of assembled chromosomes	31	32	31
Linkage group N50 (Mb)	23.78	11.37	10.31
BUSCO genes (%)	97.3	99.4	99.6
Repeat (%)	58.9	43.6	42.3
G + C (%)	39.1	36.0	35.5
Number of genes	15,855	13,876	13,255

Table 3. Summary of genome assembly statistics for *E. obscuralis*.

read pairs. In parallel, Illumina short reads were used to polish the draft assembly in two rounds using Pilon[10] (v1.23). The resulting draft genome was 701.8 Mb in size, composed of 685 contigs with a contig N50 of 1.73 Mb (Table 3).

To further organize the genome into chromosome-scale scaffolds, Hi-C reads (filtered using the same quality control procedures as described for Illumina data) were mapped to the draft genome using BWA[11] (v0.6.2). Clean paired-end reads uniquely mapped near Hi-C restriction sites were selected for downstream scaffolding. Contigs were clustered, ordered, and oriented into chromosomes using ALLHiC[12], and further manual refinement was performed with Juicebox[13]. As a result, 643 contigs—representing 96.98% of the draft genome—were anchored onto 31 chromosomes, yielding a scaffold N50 of 23.7 Mb (Fig. 1, Table 2).

The completeness of the assembly was assessed using BUSCO[14] (Benchmarking Universal Single-Copy Orthologs; http://busco.ezlab.org/), which identified 96.7% of expected orthologs, including 87.3% as single-copy genes and 9.4% as duplicates. These results demonstrate that the genome assembly is both highly complete and suitable for downstream analyses.

While the *E. obscuralis* genome presented here constitutes a high-quality, chromosome-scale assembly, several limitations should be acknowledged. First, the use of genomic DNA pooled from multiple individuals may have increased the apparent heterozygosity and introduced allelic variation, which can complicate genome assembly, particularly in repetitive or low-complexity regions. Although k-mer analysis estimated heterozygosity at 1.87%, this value likely reflects population-level diversity rather than the genetic variation within a single individual. Second, despite the integration of PacBio HiFi long reads and Hi-C chromatin interaction data, gaps persist in the assembly, primarily due to the presence of highly repetitive sequences and structurally complex regions that remain challenging to resolve with current sequencing technologies. Additionally, as DNA was extracted from whole-body samples rather than specific tissues, the presence of mitochondrial DNA and potential contamination from endosymbionts or gut content cannot be entirely excluded, despite efforts to filter such sequences during assembly and annotation. Future improvements—such as single-individual sequencing and the incorporation of ultra-long-read technologies—are expected to further enhance the contiguity and completeness of the *E. obscuralis* genome.

Sex chromosome identification. Genome-wide synteny between *E. obscuralis* and two related species—*Amyelois transitella* (accession: GCF_032362555.1; karyotype: 30 + ZW) and *Plodia interpunctella* (accession: GCF_027563975.2; karyotype: 30 + ZW)—was assessed using **Satsuma**[15] (v2.0). A collinearity plot generated with **Circos**[16] (v0.69) revealed strong syntenic relationships, particularly between *E. obscuralis* chromosome 9 and the Z chromosomes of both *A. transitella* and *P. interpunctella*, indicating conserved sex chromosome architecture among these Lepidoptera species (Fig. 2).

Genome annotation. The genome of *E. obscuralis* was annotated using a comprehensive, multi-step strategy, beginning with the identification of repetitive sequences. Repeats were detected using the HiTE[17] (High-throughput Transposable Element) pipeline, which integrates both homology-based and de novo approaches to identify transposable elements (TEs) and other repetitive elements within the genome. Specifically, HiTE first aligns the genome to known repeat databases and then applies de novo methods based on sequence composition and structural features to detect novel elements. This combined strategy enabled the annotation of approximately 58.90% of the *E. obscuralis* genome as repetitive sequences. Among these, short interspersed nuclear elements (SINEs) accounted for 1.43%, long interspersed nuclear elements (LINEs) for 9.77%, long terminal repeats (LTRs) for 14.47%, and DNA transposons for 21.29%. An additional 4.64% of repeat elements were categorized as unclassified (Table 4).

Following repeat masking, gene prediction was conducted using eGAPX (v0.3.2), a high-accuracy gene annotation tool that integrates ab initio prediction with RNA-Seq evidence. This hybrid approach enhances gene model accuracy, particularly in complex genomes. RNA-Seq data from a single female *E. obscuralis* adult were

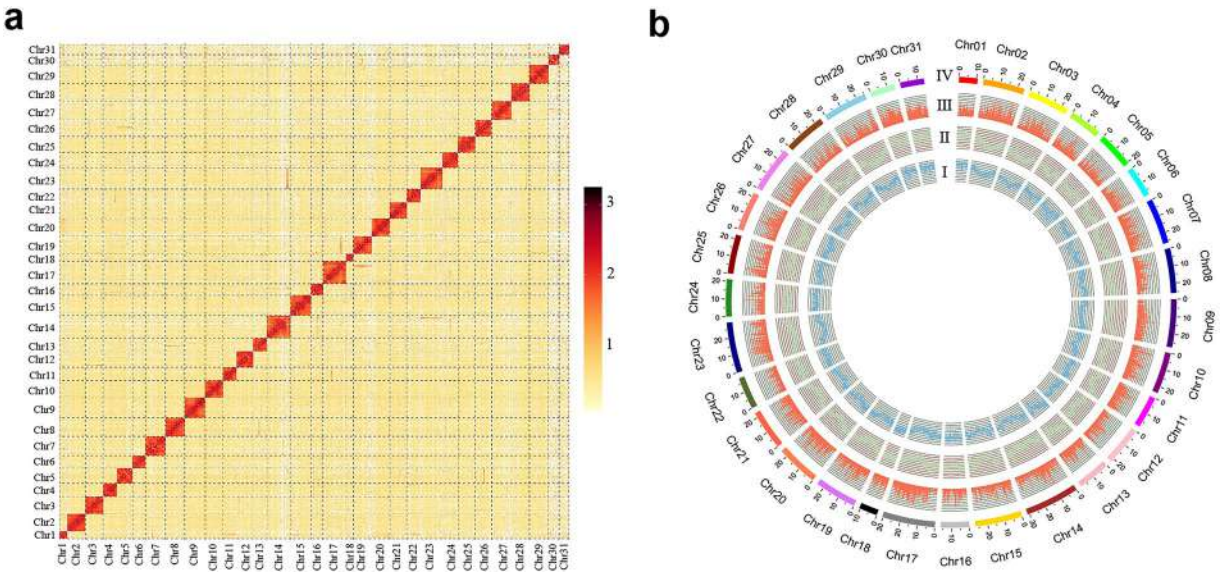

Fig. 1 Hi-C contact heatmap and genome landscape of the leaf roller (*Eumorphobotys obscuralis*). (**a**) Hi-C contact heatmap showing chromatin interaction frequencies across the nine assembled chromosomes of *E. obscuralis*. Strong intra-chromosomal interactions appear as prominent blocks along the diagonal, while weaker inter-chromosomal interactions are visible off the diagonal. Color intensity reflects the frequency of Hi-C contacts. (**b**) Circular representation of the *E. obscuralis* genome landscape using a 100 Kb sliding window. From the outermost to innermost circles: IV. Chromosome ideograms; III. Protein-coding gene density; II. GC content density; I. Repeat element density.

incorporated to refine gene boundaries and improve the annotation of exons, introns, and untranslated regions (UTRs). As a result, a total of 15,855 protein-coding genes and 19,196 transcripts were predicted. Functional annotation of these gene models was performed using BLASTP searches against several public protein databases, including NCBI's NR and UniProt Swiss-Prot. Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway assignments were also applied to predict gene functions, biological processes, and molecular interactions.

In addition to protein-coding genes, three classes of non-coding RNAs (ncRNAs)—transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), and small nuclear RNAs (snRNAs)—were systematically annotated after excluding protein-coding genes and repetitive elements. tRNA genes were identified using tRNAscan-SE[18] (v2.0) with parameters optimized for eukaryotic genomes. rRNA genes were predicted via BLASTN by aligning the genome against a curated invertebrate rRNA reference database, applying an E-value threshold of $<1 \times 10^{-5}$. snRNA genes were annotated using INFERNAL[19] in combination with the Rfam database[20] (release 14.9) to detect conserved RNA motifs. The integrated analysis identified 114 small nucleolar RNAs (snoRNAs), 20,780 tRNA genes, 60 rRNA genes, and 76 microRNA (miRNA) genes, as summarized in Table 5.

Data Records

The sequencing data generated for this study have been deposited in public repositories. The Illumina, PacBio, and Hi-C datasets used for genome assembly are available in the NCBI Sequence Read Archive (SRA) under accession numbers SRR32383175[21], SRR32383174[22], and SRR32383173[23], respectively, under the BioProject accession PRJNA1207245[24]. The transcriptome sequencing data used for genome annotation are also available in the SRA under accession number SRR32383172[25]. The chromosome-scale genome assembly has been deposited in GenBank under accession JBKJAC000000000 and is associated with the GenBank assembly accession GCA_048418795.1[26]. In addition, the assembled genome and annotated gene sets have been made publicly accessible via the FigShare repository[27].

Technical Validation

Validation analyses were conducted to evaluate the contiguity, accuracy, and completeness of the *E. obscuralis* genome assembly. The assembled genome measures 701.9 Mb, with a scaffold N50 of 23.7 Mb, closely matching the genome size estimated by k-mer analysis. The Hi-C contact heatmap revealed a well-organized interaction pattern along the diagonals, consistent with correct chromosomal assembly, and also showed clear interaction signals near potential chromosomal inversions (Fig. 1a). In addition, strong synteny was observed between *E. obscuralis* and two related moth species, supporting the structural accuracy of the chromosome-level assembly (Fig. 2).

To further assess assembly quality, Illumina short reads were mapped back to the final genome assembly, resulting in an average mapping rate of 95.58% and genome coverage of 96.50%. The completeness of the assembly was evaluated using BUSCO and CEGMA. BUSCO analysis based on the Arthropoda dataset identified

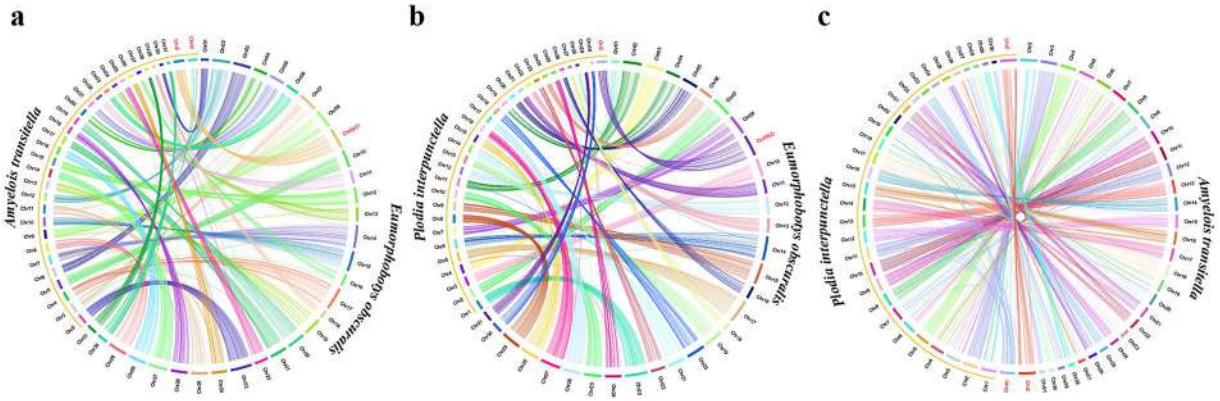

Fig. 2 Genome-wide collinearity analysis of different species by the Satusam software. (**a**) Collinearity between *E. obscuralis* and *Amyelois transitella*. (**b**) Collinearity between *E. obscuralis* and *Plodia interpunctella*. (**c**) Collinearity between *A. transitella* and *P. interpunctella*.

Category	Number of elements	Ratio (%) in genome
SINEs	82,385	1.43%
LINEs	666,102	9.77%
LTR elements	1,050,339	14.47%
DNA elements	1,122,844	21.29%
Unclassified	312,004	4.64%

Table 4. Classification and composition of repetitive sequences in the *E. obscuralis* genome.

Category		Number	Total length (bp)	Average length (bp)	% of genome
miRNA		76	5,894	77.55	0.000008397
tRNA		20,780	1,548,148	74.50	0.002205498
rRNA	18S	8	10,747	1343.38	0.000015310
	28S	9	16,661	1851.22	0.000023735
	5.8S	4	628	157.00	0.000000895
	5S	39	4,620	118.46	0.000006582
snRNA	CD-box	18	2,546	141.44	0.000003627
	HACA-box	7	893	127.57	0.000001272
	splicing	89	12,533	140.82	0.000017855

Table 5. Classification and distribution of non-coding RNAs in the *E. obscuralis* genome.

96.7% of expected genes, including 87.3% as complete and single-copy and 9.4% as duplicated. CEGMA evaluation using the eukaryotic core gene dataset reported 82.66% completeness.

Functional validation was also performed by aligning annotated genes against several public protein databases. 80.42% of genes matched entries in the NCBI RefSeq database. Matches were also found in EggNOG (76.28%), Swiss-Prot (67.05%), and KOG (49.79%), further confirming the reliability and completeness of the gene annotation.

Code availability

Programs used in data processing were executed with default parameters unless otherwise specified in the Methods section. No custom code was employed for these analyses.

Received: 24 March 2025; Accepted: 16 June 2025;
Published online: 23 June 2025

References

1. Caradja, A. Ueber Chinas Pyraliden, Tortriciden, Tineiden nebst kurzen Betrachtungen, zu denen. das Studium dieser Fauna Veranlassung gibt (Eine biogeographische Skizze). *Memle Sect. Stiint. Acad. rom.* **3**, 257–383 (1925).
2. Dou, Y., Yu, X. & Fumiyo, I. The current situation and countermeasures of bamboo resource development and utilization of China. *Chinese Journal of Agricultural Resources and Regional Planning* **32**, 65–70 (2011).

3. Scheffers, M. F., Ona Ayala, K. E., Ottesen, T. D. & Tuakli-Wosornu, Y. A. Design and development of mobility equipment for persons with disabilities in low-resource and tropical settings: bamboo wheelchairs. *Disabil Rehabil-Assist* **16**, 377–383 (2021).
4. Ying, Y. *et al*. Mitochondrial Genome Comparison and Phylogenetic Variety of Four Morphologically Similar Bamboo Pests. *Ecol Evol* **14**, e70588 (2024).
5. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
6. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
7. Liu, B. *et al*. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. arXiv preprint *arXiv:1308.2012* (2013).
8. 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).
9. Wingett, S. *et al*. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research* **4** (2015).
10. Walker, B. J. *et al*. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one* **9**, e112963 (2014).
11. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics* **25**, 1754–1760 (2009).
12. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat Plants* **5**, 833–845 (2019).
13. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst* **3**, 99–101 (2016).
14. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness. *Gene prediction: methods and protocols*, 227-245 (2019).
15. Grabherr, M. G. *et al*. Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics* **26**, 1145–1151 (2010).
16. Krzywinski, M. *et al*. Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639–1645 (2009).
17. Hu, K. *et al*. HiTE: a fast and accurate dynamic boundary adjustment approach for full-length transposable element detection and annotation. *Nat. Commun* **15**, 5573 (2024).
18. 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).
19. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
20. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S. R. Rfam: an RNA family database. *Nucleic Acids Res* **31**, 439–441 (2003).
21. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR32383175 (2025).
22. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR32383174 (2025).
23. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR32383173 (2025).
24. *NCBI BioProject* https://identifers.org/ncbi/bioproject:PRJNA1207245 (2025).
25. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR32383172 (2025).
26. *NCBI Genome* https://identifiers.org/insdc.gca:GCA_048418795.1 (2025).
27. Dong, Y. M. *et al*. The genome annotation of the leaf roller, *Eumorphobotys obscuralis* (Lepidoptera: Pyralidae). *Figshare. Collection.* https://doi.org/10.6084/m9.figshare.28546304.v1 (2025).

Acknowledgements

This work was supported by the National Science Foundation of China (No. 32172402), the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2023C02034), and the Key Investigation and Monitoring of Agricultural Alien Invasive Species Project of the Ministry of Agriculture and Rural Affairs of China (No. No.019240117).

Author contributions

J.M.Z., J.P.S. and P.J.Z. designed the project. S.Z.W., J.R.Z. and W.Z. coordinated the study. Y.Y. conducted the sampling and sequencing; P.J.Z. analyzed the genome size; J.M.Z. annotated the genome; J.P.S. performed the chromosomal syntheses analysis, comparative genomics analysis, and gene family identification; S.Z.W. and J.R.Z. drafted the manuscript, and P.J.Z. improved and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.-P.S., J.-M.Z. or P.-J.Z.