scientific data

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of the teak defoliator *Hyblaea puera* Cramer (Lepidoptera: Hyblaeidae)

Zixuanxin Wang[1,4], Dezhi Kong[1,2,4], Yunpeng Liu[1], Tonghai Zhao[1,2], Qinghua Wang[1,2], Jicheng Zhao[1], Yaojun Zhu[3] & Liangjian Qu[1,2] ✉

The teak defoliator, *Hyblaea puera*, is a major pest damaging teak plantations and mangrove ecosystems. The high reproductive and dispersal capacities of this pest often lead to sudden and explosive outbreaks, posing challenges for its control. Genomic insights into this pest could help to elucidate its ecological adaptation and inform effective management strategies. Here, a high-quality genome of *H. puera* was assembled using integrated Illumina, PacBio HiFi, and Hi-C sequencing data. The final chromosome assembly is 394.93 Mb with a contig N50 of 11.49 Mb. A total of 96.97% of assembled sequences were anchored to 31 pseudo-chromosomes. Approximately 26.64% of the genome sequences comprise repetitive elements. Genome annotation identified 15,364 protein-coding genes, of which 86.79% (13,334) were functionally characterized. The completeness of the assembled genome was estimated to be 97.73% using BUSCO. This high-quality genome provides valuable resources for understanding the ecological adaptation and developing control strategies of *H. puera*.

Background & Summary

The teak defoliator *Hyblaea puera* Cramer (Lepidoptera: Hyblaeidae) is a destructive tropical lepidopteran pest. The larvae feed on leaves and young shoots, impairing tree growth and potentially causing tree mortality in severe cases[1–3]. This pest is native to South and Southeast Asia, including India, Laos, and Myanmar, but has now spread to at least 34 countries across Asia, Oceania, Central America, and Africa through international trade and plant introductions[3–5]. *H. puera* primarily feeds on teak (*Tectona grandis*) and mangrove species. In China, researchers first documented its presence in teak plantations in 1975[1]. By 2010, infestations had been recorded in the mangrove areas of Guangxi Province, particularly affecting *Avicennia marina*[6]. Since 2015, recurrent large-scale outbreaks of *H. puera* have been reported in mangrove regions across Guangxi and Guangdong provinces, posing an increasing threat to the stability and coastal protection functions of Chinese mangrove ecosystems[5–8].

The outbreak of *H. puera* is associated with its ecological adaptation. First, the *H. puera* demonstrates a remarkable reproductive capacity, which is a key driver of outbreaks[6]. Second, adult moths typically undertake short-range dispersal (10–20 km) to find suitable host plants when local foliage is exhausted or becomes nutritionally suboptimal[9]. These features cause the pest's explosive and unpredictable outbreak patterns, posing major challenges for its control[5]. Genomic research on this species could help uncover the genetic basis of its ecological adaptability and provide valuable insights for effective management strategies. However, the lack of a reference genome restricts current research on *H. puera*[2,4,10].

In the present study, whole-genome sequencing of *H. puera* was performed using female adult specimens. The genome size was initially estimated by k-mer analysis (k = 19) based on Illumina data. The Illumina sequencing obtained 32.55 Gb (Table 1) of clean data, which predicted a genome size of 338.39 Mb (Fig. 1), along

[1]Key Laboratory of Forest Protection of National Forestry and Grassland Administration, Ecology and Nature Conservation Institute, Chinese Academy of Forestry, Beijing, 100091, P. R. China. [2]Co-Innovation Center for Sustainable Forestry in Southern China, Nanjing Forestry University, Nanjing, 210037, P. R. China. [3]Zhanjiang National Research Station for Mangrove Wetland Ecosystem, Zhanjiang, 524448, P. R. China. [4]These authors contributed equally: Zixuanxin Wang, Dezhi Kong. ✉e-mail: qulj2001@caf.ac.cn

Sequencing strategy	Platform	Usage	Insertion size	Clean data (Gb)	Coverage (×)
Short reads	Illumina	Genome survey	350 bp	32.55	96
Long reads	PacBio HiFi	Assembly	15 kb	15.20	44
Hi-C	Illumina	Hi-C assembly	300 bp	40.61	106
RNA-seq	Illumina	Annotation evidence	350 bp	7.24	20

Table 1. Sequencing methods and the generated data used for *Hyblaea puera* genome assembly.

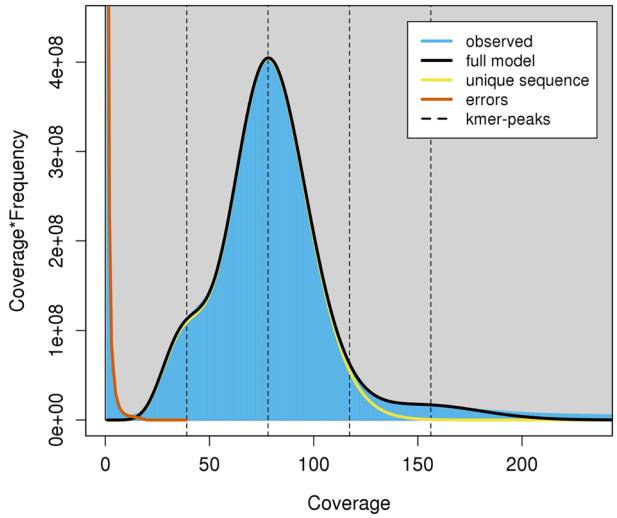

Fig. 1 The genomescope profile of the *Hyblaea puera* genome under k = 19.

Statistics		Value
Survey	Estimated genome size (Mb)	338.39
	Heterozygosity rate (%)	0.86%
	Duplication rate (%)	23%
Assembly	Total length (Mb)	394.93
	Scaffold number	148
	Longest scaffold length (Mb)	26.93
	Scaffold N50 (Mb)	12.66
	Scaffold N90 (Mb)	7.95
	GC content (%)	35.81%
	Anchored to pseudo-chromosome (%)	96.97%
Annotation	Number of annotated protein-coding genes	15,364
	Number of functionally annotated genes	13,334

Table 2. Statistics for the genome assembly of *Hyblaea puera*.

with a heterozygosity rate of 0.86% and a duplication rate of 2.76% (Table 2). Subsequently, long-read sequencing was conducted on the PacBio Sequel II platform, generating 15.20 Gb (Table 1) of HiFi data with an average read length of 14.75 kb and an N50 of 17.56 kb. To facilitate chromosome-level scaffolding, a Hi-C library was constructed and sequenced, yielding 40.61 Gb (approximately 106 × coverage) of clean data (Table 1).

The integrated assembly process resulted in a final genome size of 394.93 Mb with a GC content of 35.81%, consisting of 148 scaffolds (scaffold N50 = 12.66 Mb) (Table 2). Of these, 96.97% (382.97 Mb) of the assembled sequences were anchored into 31 pseudo-chromosomes, with lengths ranging from 5.63 Mb to 26.93 Mb (Fig. 2). This result is highly consistent with our karyotype analysis (2n = 60) (Fig. S1). To evaluate assembly quality, BUSCO assessment indicated a completeness of 97.73%, confirming the high continuity and reliability of the genome. Moreover, a comparative analysis of assembly quality with the seven lepidopteran insect reference genomes revealed that the contig N50 of *H. puera* is significantly longer than those of its close relatives (Fig. 3), further supporting the superior contiguity of the assembly.

Analysis revealed that 26.64% of the genome was annotated as repeat sequences. Among these, DNA transposons (14.23%), long interspersed nuclear elements (LINEs, 4.65%), Gypsy (3.46%) and short interspersed nuclear elements (SINEs, 1.50%) represented the four most abundant types (Table 3). Additionally, 5.58% of the

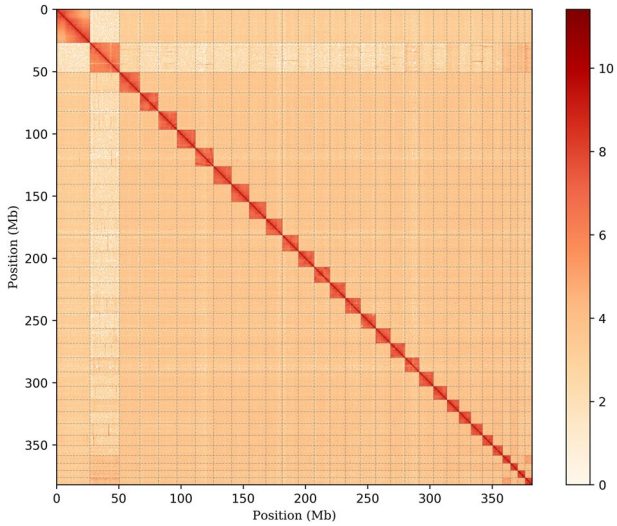

Fig. 2 Genome-wide contact matrix of *Hyblaea puera* generated by Hi-C data. Each square represents a pseudo-chromosome.

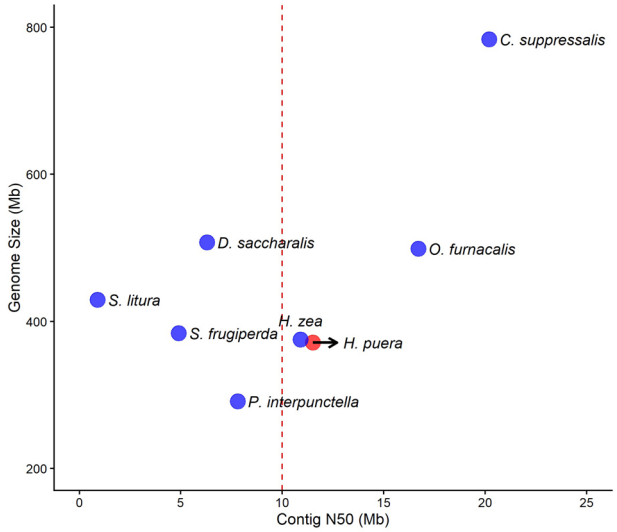

Fig. 3 Scatter plot showing the contig N50 distribution of genome assemblies for *Hyblaea puera* and its relative species.

repeat sequences in the genome remained unclassified. Regarding protein-coding genes, a total of 15,364 genes were predicted in the assembled genome. The landscape of *H. puera* genome assembly and gene annotations was presented as a circos plot (Fig. 4). The functions of protein-coding genes were annotated against multiple databases. The results suggested that 86.79% (13,334) of coding genes could be functionally annotated by at least one public database and transcriptome (Table 4). Furthermore, various types of non-coding RNAs were identified, including 8,862 tRNA genes, 502 rRNA loci, 71 snRNA elements, and 60 miRNA genes.

Orthologous genes were identified by comparing *H. puera* with 14 other insect species, leading to the identification of 11,227 gene families in this insect. To elucidate evolutionary relationships, a phylogenetic tree was constructed using the protein sequences of 2,034 single-copy orthologous genes identified by OrthoFinder (Table 5). The analysis revealed that *H. puera* forms a distinct branch among the species examined, with an estimated divergence time of approximately 110.23 million years ago (MYA) (Fig. 5). Furthermore, gene family expansion and contraction were analyzed using CAFE (v4.2.1), based on the species phylogeny with divergence times and the orthogroups inferred by OrthoFinder. A total of 103 gene families were identified as significantly expanded and 4 as significantly contracted in *H. puera* (Fig. 5), reflecting significant changes in gene number during its evolution.

Taken together, a high-quality chromosome-level genome was assembled for *H. puera*, using a combination of Illumina, HiFi, and Hi-C sequencing data. This genomic resource provides fundamental data for investigating the ecological adaptations, evolutionary mechanisms, and pest management of *H. puera*.

Class	Number	Length	Percentage (%)
LTR			
Copia	5,643	2,088,752	0.53
Gypsy	25,571	13,647,348	3.46
Others	52,211	8,902,167	2.25
Non-LTR retroelements			
LINE	71,351	18,371,253	4.65
SINE	33,218	5,922,143	1.50
DIRS	218	46,082	0.01
DNA transposons	353,382	56,191,208	14.23
Satellite	128,665	6,766,537	1.71
Unclassified	114,224	22,046,851	5.58
Unknown	200	25,914	0.01
Total		105,194,867	26.64

Table 3. Annotation of repeat elements in the genome of *Hyblaea puera*.

Methods

Sample preparation. The pupae of *H. puera* were collected in Beihai, Guangxi Zhuang Autonomous Region, China, in August 2023. The larvae were reared in the laboratory for over 20 generations on an artificial diet under controlled conditions: a temperature of 29 ± 1 °C, a 16 L:8 D photoperiod, and a relative humidity of 70 ± 5%. Genomic DNA was extracted from two adult females, and RNA for transcriptome sequencing was obtained from two additional adults (one male and one female). To minimize potential microbial contamination, the guts and mouthparts of the adults were removed. The prepared samples were immediately flash-frozen in liquid nitrogen and stored at −80 °C until analysis. For Hi-C sequencing, the tissue of one adult female was fixed with formaldehyde prior to DNA extraction.

Genomic DNA and RNA sequencing. For short-read sequencing, the genomic DNA was isolated from the female adults using the TGuide Smart Universal DNA Kit (Tiangen Biotech). A 350 bp insert-size paired-end library was constructed with the NEB DNA Library Rapid Prep Kit (NEB, MA, USA). Sequencing was performed on an Illumina NovaSeq X system (Illumina, San Diego, CA, USA).

For long-read sequencing, the genomic DNA from female adults was first subjected to fragmentation and end polishing, followed by adapter ligation and size selection to generate sequencing-ready fragments. The SMRTbell library was prepared using the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences) according to the manufacturer's protocol, with purification performed using 0.45 × AMPure PB beads. Library quality control was carried out using two complementary approaches: fragment size distribution was analyzed on a FEMTO Pulse capillary electrophoresis system (Agilent Technologies), while DNA concentration was precisely quantified using a Qubit 3.0 Fluorometer (Thermo Fisher Scientific). The qualified library was subsequently sequenced on a PacBio SequelII platform.

For Hi-C sequencing, fresh tissues from the female adults were crosslinked with formaldehyde to fix cellular structures, thereby stabilizing protein-DNA and DNA-DNA interactions and preserving the native three-dimensional architecture. Crosslinked genomic DNA was subsequently digested with the restriction enzyme HindIII to generate DNA fragments with sticky ends. An end-repair reaction was then performed, incorporating biotin-labeled nucleotides to provide specific tags for downstream steps[11]. The repaired DNA fragments underwent proximity ligation-mediated circularization, ensuring the covalent linkage of spatially interacting DNA fragments. Following circularization, crosslinks were reversed to release the DNA, which was then purified and randomly sheared. Fragments ranging from 300 to 700 bp were selected, with the peak fragment size centered at 300 bp. Biotin-labeled DNA fragments, representing interaction junctions, were specifically enriched using streptavidin-coated magnetic beads. Finally, a high-throughput sequencing-ready Hi-C library was constructed. Post-construction, library concentration and insert size were assessed using Qubit 2.0 (Thermo Fisher Scientific, USA) and Agilent 2100 (Agilent Technologies, Santa Clara, CA), respectively. The library was sequenced on an Illumina NovaSeq X platform using a paired-end 150-bp (PE150) strategy.

For transcriptome sequencing, total RNA was extracted from female adults using TRIzol reagent (Thermo Fisher Scientific, USA). Paired-end libraries were prepared with the Hieff NGS Ultima Dual-mode mRNA Library Prep Kit for Illumina (Yeasen Biotechnology, Shanghai, China) in accordance with the manufacturer's instructions. The resulting libraries were sequenced on an Illumina NovaSeq X platform with a paired-end 150-bp (PE150) strategy.

Estimation of genomic characteristics. Quality control of the short reads from the Illumina NovaSeq X platform was performed using fastp[12], following the manufacturer's recommended protocols. The genome size of *H. puera* was estimated by the k-mer method using Illumina clean data. K-mer distribution was estimated using Jellyfish v2.1.4[13] with parameters -m 19. The heterozygosity rate was estimated by GenomeScope v2.0[14].

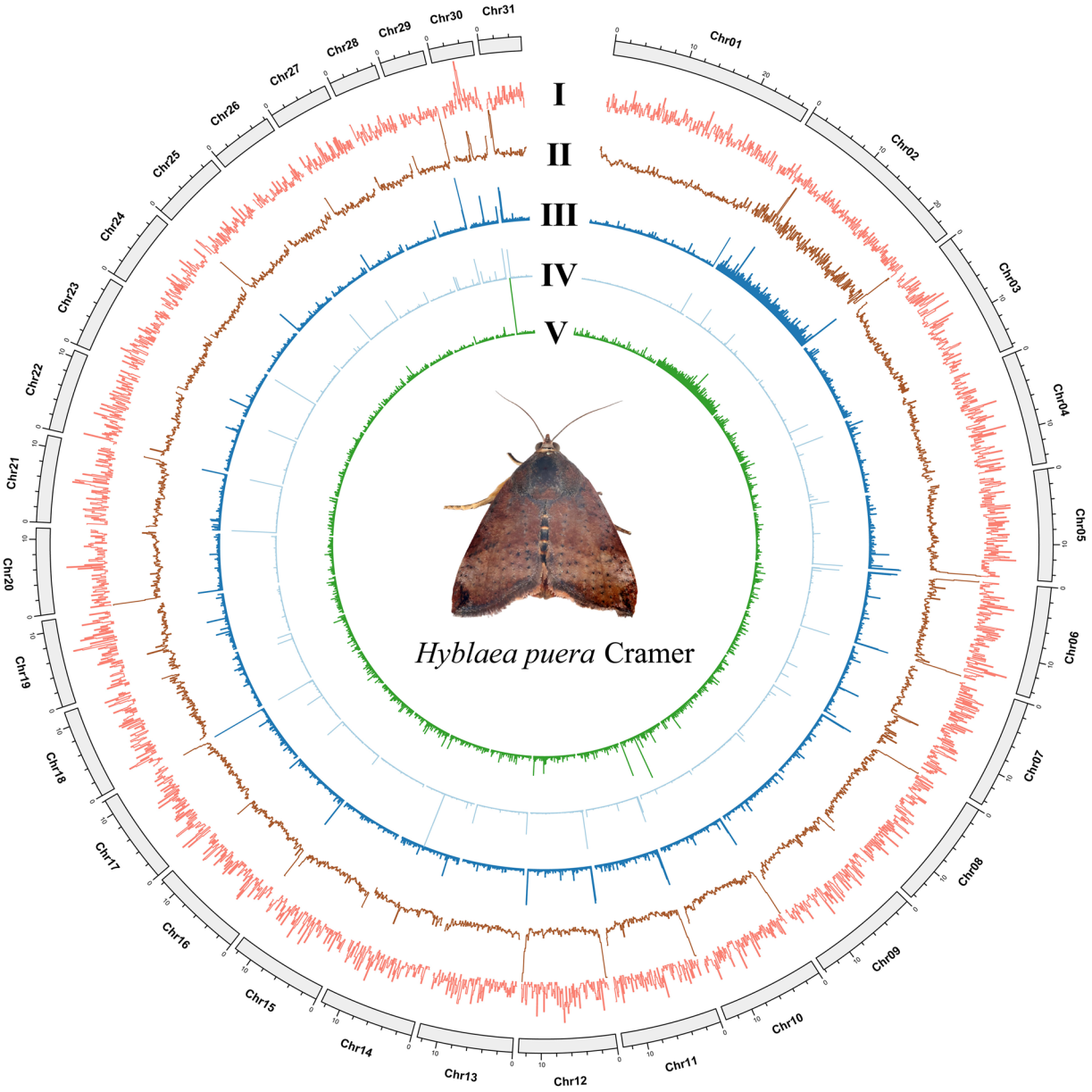

Fig. 4 Overview of the genome landscape of *Hyblaea puera*. In the circos plot, the outermost track denotes the pseudo-chromosomal ideograms (scale = 1 Mb). From the outer to the inner tracks, the density of protein-coding genes (I), GC content (II), TE density (III), SSR density (IV), and tRNA density (V) on each pseudo-chromosome was calculated in nonoverlapping 100-kb windows and displayed.

Database	Number of annotated genes	Ratio (%)
GO	10784	70.19
KEGG	11643	75.78
KOG	9234	60.10
Pfam	12177	79.26
SwissProt	7321	47.65
TrEMBL	11084	72.14
eggNOG	9674	62.97
nr	10326	67.21
All	13334	86.79

Table 4. Summary of gene function annotation.

Number of genes	210532
Number of genes in orthogroups	190031
Number of unassigned genes	20501
Percentage of genes in orthogroups	90.3
Percentage of unassigned genes	9.7
Number of orthogroups	19891
Number of species-specific orthogroups	1655
Number of genes in species-specific orthogroups	5902
Percentage of genes in species-specific orthogroups	2.8
Mean orthogroup size	9.6
Median orthogroup size	8.0
G50 (assigned genes)	14
G50 (all genes)	14
O50 (assigned genes)	5205
O50 (all genes)	5937
Number of orthogroups with all species present	2844
Number of single-copy orthogroups	2034

Table 5. Summary of the orthogroups across 15 insect species in this study.

Genome assembly and quality evaluation. The *H. puera* genome was initially assembled using PacBio HiFi reads, which were processed with Hifiasm v0.19[15]. Contaminant sequences were identified and removed through comprehensive alignment against the NCBI NT, mitochondrial and plastid reference databases. The resulting draft assembly comprised 301 contigs with a total length of 468.66 Mb, exhibiting a contig N50 of 11.49 Mb, the longest contig of 25.93 Mb, and an overall GC content of 35.75%. To achieve a chromosome-scale genome, Hi-C scaffolding was performed, where the Hi-C reads were first aligned to the draft assembly using BWA v0.7.17[16] with default parameters and then processed through Juicer v1.6[17] and 3D-DNA[18] to sort, orient, and order contigs into chromosomal scaffolds.

Additionally, to investigate the quality of the assembled genome, a comparison of contig sizes between *H. puera* and related species[10] was carried out. Based on phylogenetic analysis of lepidopteran mitochondrial genome sequences, seven insect species with sequenced genomes that show a close phylogenetic relationship with *H. puera* were selected for contig size comparison.

Repeat sequences and non-coding RNA annotation. Transposable elements (TEs) and tandem repeats were identified by a combination of a homology-based method and a *de novo* approach. The initial steps involved generating a *de novo* repeat library through the synergistic application of RepeatModeler[19] and LTR_retriever v2.9.0[20]. The latter tool processed and refined long terminal repeat (LTR) predictions derived from parallel analyses with LTRharvest v1.5.9[21] and LTR_finder v1.0.7[22]. Subsequent integration with the Dfam database produced a non-redundant, species-specific TE repository. Genome-wide TE identification and classification used RepeatMasker v4.12[23] to perform homology searches against a customized library (Table 2). Concurrent tandem repeat analysis implemented complementary approaches using Tandem Repeats Finder v4.09[24] and the MISA v2.1[25] microsatellite detection system.

The non-coding RNAs (ncRNAs) were characterized through a multi-step bioinformatics analysis. Initial identification of ncRNA elements was performed through whole-genome alignment against the Rfam database v14.5[26], followed by specialized prediction of distinct ncRNA classes: tRNAscan-SE v2.0.12[27] was utilized for transfer RNA identification, Barrnap v0.9 for ribosomal RNA prediction, and Infernal v1.1.5[28] with covariance models for both small nucleolar RNAs and microRNAs.

Protein-coding gene prediction and functional annotation. Protein-coding genes in the genome were predicted using a combination of three complementary methods: *de novo* prediction, homology-based search, and transcriptome-guided assembly. *De novo* gene prediction was carried out using two ab initio tools, AUGUSTUS v3.1.0[29] and SNAP (2006-07-28). For homology-based gene identification, GeMoMa v1.7[30] was employed, using reference genomes of *Chilo suppressalis*[31], *Diatraea saccharalis*[32], *Ostrinia furnacalis*[33], and *Plodia interpunctella*[34]. Transcriptomic evidence was incorporated by mapping RNA-seq reads to the reference genomes with HISAT2 v2.1.0[35], followed by transcript assembly using StringTie v2.1.4[36]. GeneMarkS-T v5.1[37] was subsequently applied to predict genes from the assembled transcripts. Additionally, full-length transcripts and unigenes generated by Trinity v2.11[38] were processed through PASA v2.4.1[39] to improve the prediction. Gene predictions from all three approaches were integrated using EVidenceModeler (EVM, v1.1.1)[40] and further polished with PASA.

The function of predicted genes was annotated using BLASTP (E-value < 1e-05) against databases including Eukaryotic Orthologous Groups (KOG), eggNOG[41], SwissProt[42], TrEMBL[42], and the NCBI non-redundant proteins database (NR) using DIAMOND blastp[43]. The structural domains and motifs of all genes were scanned against Pfam databases using InterProScan v5.34-73.0[44]. The Gene Ontology (GO) terms were extracted based on the corresponding InterPro entry. The metabolic pathways in which the genes might be involved were

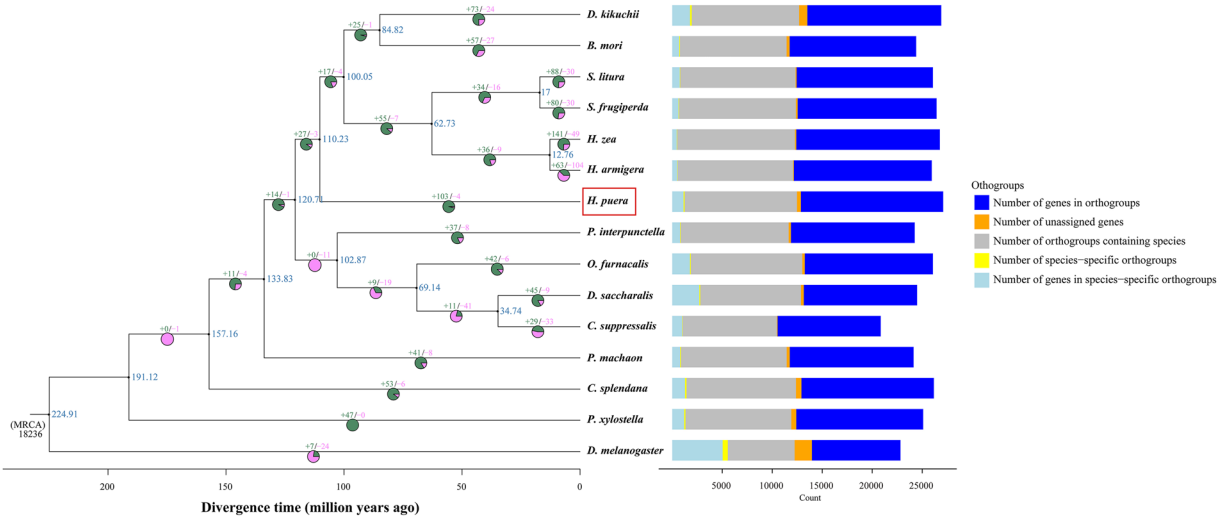

Fig. 5 Phylogenetic tree and gene ortholog analysis of 15 insect species. The maximum likelihood tree includes *Hyblaea puera* (Lepidoptera) and 13 other lepidopteran species, with *Drosophila melanogaster* included as an outgroup. Divergence times are indicated at the nodes throughout the phylogeny. Gene family expansions and contractions inferred from CAFE analysis are highlighted in pink and green, respectively. The bar plot on the right displays the distribution of orthologous gene counts for each species.

assigned by BLAST (E-value < 1e-05) against the Kyoto Encyclopedia of Genes and Genomes (KEGG) protein database[45].

Identification of orthologs and inference of phylogenetic relationships. To establish a reliable phylogenetic framework, this study conducted a comparative genomic analysis of *H. puera* and 14 related species. The analyzed taxa included *Dendrolimus kikuchii*[46], *Bombyx mori*[47], *Spodoptera litura*, *Spodoptera frugiperda*, *Helicoverpa zea*, *Helicoverpa armigera*, *Ectomyelois gemmivora*, *Plodia interpunctella*, *Ostrinia furnacalis*, *Diatraea saccharalis*, *Chilo suppressalis, Papilio machaon*[48], *Cydia splendana* and *Plutella xylostella*[49], with *Drosophila melanogaster* designated as the outgroup. All genomic data and annotated protein sequences were obtained from public databases such as NCBI. For each gene, only the longest transcript isoform was retained for subsequent analysis (Fig. 5).

OrthoFinder (v2.4)[50] with DIAMOND (E-value threshold: 0.001) was employed to identify orthologous groups, resulting in the identification of 2,034 single-copy orthologous genes. For phylogenetic tree construction, multiple sequence alignment for each single-copy gene was performed using MAFFT (v7.205)[51]. The aligned sequences were then refined with Gblocks (v0.91b)[52] to remove poorly aligned regions, and the filtered alignments for each species were concatenated into a supergene sequence. Phylogenetic trees were constructed using IQ-TREE (v1.6.11)[53]. The best-fit amino acid substitution model (LG + F + I + G4) was first determined by the built-in ModelFinder[54] algorithm. Then, maximum likelihood trees were inferred under this model with 1,000 bootstrap replicates to assess node support. Divergence times were estimated based on the maximum likelihood phylogeny using the MCMCtree program (PAML v4.9i)[55], with calibrations derived from the TimeTree database (Fig. 5).

Gene family expansion and contraction. The phylogenetic tree with divergence times and the gene family clustering results were used as inputs to infer significant gene family expansion and contraction using CAFE v4.2[56], with a significance threshold of a family-wide P-value < 0.05 and a Viterbi P-value < 0.05.

Data Record

The genomic data associated with this study are available under NCBI BioProject PRJNA1289346[57]. The short-reads, long-reads, Hi-C, and RNA-seq data are available in the NCBI Sequence Read Archive under the accession numbers of SRR34562996[58], SRR34567164[59], SRR34577075[60], and SRR34587760[61], respectively. The chromosome-level genome assembly has been submitted to NCBI GenBank with the accession number JBPQDY020000000[62]. The genome assembly and annotation data have been deposited in the Figshare database[63].

Technical Validation

To assess the completeness and uniformity of the genome assembly, the Illumina short reads and PacBio HiFi long reads were mapped back to the assembly using BWA and Minimap2[64], respectively. The mapping rates were 99.31% for the short reads and 99.93% for the long reads, indicating a high-quality and continuous assembly.

The completeness of the genome assembly and annotation was evaluated using Benchmarking Universal Single-Copy Orthologs (BUSCO v5.2.2) with the insecta_odb10 database (1,367 conserved genes)[65]. For the genome assembly, BUSCO analysis revealed a high level of completeness, with 97.73% of conserved genes identified as complete (94.88% single-copy, 2.85% duplicated), 0.07% fragmented, and 2.19% missing. For the gene

Type	Assembled genome		Annotated protein-coding genes	
	Number	Percent (%)	Number	Percent (%)
Complete gene (C)	1336	97.73	1335	97.66
Complete and single-copy gene (S)	1297	94.88	1295	94.73
Complete and duplicated gene (D)	39	2.85	40	2.93
Fragmented gene (F)	1	0.07	0	0.00
Missing gene (M)	30	2.19	32	2.34
Total	1367	100	1367	100

Table 6. Summary of BUSCO analysis of the *Hyblaea puera* genome completeness.

annotation, BUSCO results indicated even higher completeness (97.66%), including 2.93% duplicated genes. No fragmented genes were detected (0%), with only 2.34% of genes identified as missing (Table 6).

Data availability

The genome assembly and supporting raw sequence data (short-read, long-read, Hi-C, and RNA-seq) for *Hyblaea puera* have been deposited at NCBI. The project is registered under BioProject PRJNA1289346[57]. The chromosome-level assembly is accessible via GenBank accession JBPQDY020000000[62], and all raw reads are available in the Sequence Read Archive (SRA) under accessions SRR34562996[58], SRR34567164[59], SRR34577075[60], and SRR34587760[61]. Additionally, the genome assembly and its annotation files are also hosted on Figshare[63].

Code availability

All software tools employed in this study represent standard bioinformatic applications utilized according to the instructions. Detailed descriptions of software versions and specific parameters are comprehensively documented in the Methods section.

Received: 31 July 2025; Accepted: 3 March 2026;
Published online: 13 March 2026

References

1. Chen, Z. Q. & Wu, S. X. Preliminary observation of the *Hyblaea puera* (Cramer). *Trop. For.* **6**, 19–22 (1978).
2. Chandrasekhar, N., Neetha N. V., Linda Koshy Vaidyan, & Moinak Banerjee. Deciphering the molecular phylogenetics of family Hyblaeidae and inferring the phylogeographical relationships using DNA barcoding. *J. Genet. Mol. Biol.* **19**, 158–167 (2008).
3. Raja Rishi, R. & Sundararaj, R. Record of an epidemic outbreak of *Hyblaea puera* Cramer (Lepidoptera: Hyblaeidae) on *Avicennia marina* in the mangroves of Maharashtra, India. *J. Entomol. Zool. Stud.* **8**, 306–309 (2020).
4. Biji, C. P., Sudheendrakumar, V. V. & Sajeev, T. V. Quantitative estimation of *Hyblaea puera* NPV production in three larval stages of the teak defoliator, *Hyblaea puera* (Cramer). *J. Virol. Methods* **136**, 78–82, https://doi.org/10.1016/j.jviromet.2006.04.001 (2006).
5. Gao, C. B. *et al.* Beware of the outbreak of teak moth (*Hyblaea puera* Cramer), a major pest in mangroves. *Plant Quar.* **39**, 1 (2025).
6. Liu, W. A. & Li, L. F. Characteristics and control of a new pest of *Avicennia marina*— *Hyblaea puera* Cramer. *Guangxi Sci.* **24**, 523–528 (2017).
7. Li, P. The war on moth control in mangroves. *China Ocean News* **9**, 3 (2015). October 2015.
8. Hu, R., Chen, H., Yang, K. X., Cai, B. & Li, J. H. Review of a new pest, *Hyblaea puera* (Cramer), in mangrove forest in China. *For. Pest Dis.* **35**, 34–37 (2016).
9. Liu, Y. P. *et al.* Determination of the flight ability of *Hyblaea puera*. *Chin. J. Appl. Entomol.* **62**, 807–815 (2025).
10. Shah, R. A., Riyaz, M., Ignacimuthu, S. & Sivasankaran, K. Characterization of four mitochondrial genomes from superfamilies Noctuoidea and Hyblaeoidea with their phylogenetic implications. *Sci. Rep.* **12**, 18926, https://doi.org/10.1038/s41598-022-21502-y (2022).
11. Belton, J. M. *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276, https://doi.org/10.1016/j.ymeth.2012.05.001 (2012).
12. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, 884–890, https://doi.org/10.1093/bioinformatics/bty560 (2018).
13. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**(6), 764–770, https://doi.org/10.1093/bioinformatics/btr011 (2011).
14. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun* **11**, 1432, https://doi.org/10.1038/s41467-020-14998-3 (2020).
15. 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, https://doi.org/10.1038/s41592-020-01056-5 (2021).
16. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760, https://doi.org/10.1093/bioinformatics/btp324 (2009).
17. Durand, N. C. *et al.* Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst* **3**, 95–98, https://doi.org/10.1016/j.cels.2016.07.002 (2016).
18. Dudchenko, O. *et al. De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95, https://doi.org/10.1126/science.aal3327 (2017).
19. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl Acad. Sci. USA* **117**, 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
20. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422, https://doi.org/10.1104/pp.17.01310 (2018).
21. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18, https://doi.org/10.1186/1471-2105-9-18 (2008).
22. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268, https://doi.org/10.1093/nar/gkm286 (2007).
23. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. In *Current Protocols in Bioinformatics* (ed. Baxevanis, A. D.) **vol. 25**, Unit 4.10, https://doi.org/10.1002/0471250953.bi0410s25 (2009).

24. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580, https://doi.org/10.1093/nar/27.2.573 (1999).
25. Beier, S., Thiel, T., Münch, T., Scholz, U. & Mascher, M. MISA-web: a web server for microsatellite prediction. *Bioinformatics* **33**, 2583–2585, https://doi.org/10.1093/bioinformatics/btx198 (2017).
26. Griffiths-Jones, S. *et al*. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**, D121–D124, https://doi.org/10.1093/nar/gki081 (2005).
27. 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, https://doi.org/10.1093/nar/25.5.955 (1997).
28. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935, https://doi.org/10.1093/bioinformatics/btt509 (2013).
29. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, W309–W312, https://doi.org/10.1093/nar/gkh379 (2004).
30. Keilwagen, J., Hartung, F. & Grau, J. GeMoMa: Homology-based gene prediction utilizing intron position conservation and RNA-seq data. *Methods Mol. Biol.* **1962**, 161–177, https://doi.org/10.1007/978-1-4939-9173-0_9 (2019).
31. Ma, W. *et al*. A chromosome-level genome assembly reveals the genetic basis of cold tolerance in a notorious rice insect pest, Chilo suppressalis. *Mol. Ecol. Resour.* **20**, 268–282, https://doi.org/10.1111/1755-0998.13078 (2020).
32. King, R. *et al*. The complete genome assemblies of 19 insect pests of worldwide importance to agriculture. *Pestic. Biochem. Physiol.* **191**, 105339, https://doi.org/10.1016/j.pestbp.2023.105339 (2023).
33. Liu, K., Zhang, T., Zhang, Y., Wang, Z. & He, K. Chromosome-level genome of a multivoltine biotype Ostrinia furnacalis strain. *Sci. Data* **12**(1), 881, https://doi.org/10.1038/s41597-025-05252-w (2025).
34. NCBI GenBank https://identifiers.org/ncbi/insdc.gca:GCF_027563975.2.
35. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360, https://doi.org/10.1038/nmeth.3317 (2015).
36. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295, https://doi.org/10.1038/nbt.3122 (2015).
37. Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res.* **43**, e78, https://doi.org/10.1093/nar/gkv227 (2015).
38. Grabherr, M. G. *et al*. Trinity: reconstructing a full-length transcriptome without a genome from RNA-seq data. *Nat. Biotechnol.* **29**, 644–652, https://doi.org/10.1038/nbt.1883 (2011).
39. Haas, B. J. *et al*. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666, https://doi.org/10.1093/nar/gkg770 (2003).
40. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7, https://doi.org/10.1186/gb-2008-9-1-r7 (2008).
41. Huerta-Cepas, J. *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314, https://doi.org/10.1093/nar/gky1085 (2019).
42. Boeckmann, B. *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370, https://doi.org/10.1093/nar/gkg095 (2003).
43. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410, https://doi.org/10.1016/S0022-2836(05)80360-2 (1990).
44. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240, https://doi.org/10.1093/bioinformatics/btu031 (2014).
45. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109–D114, https://doi.org/10.1093/nar/gkr988 (2012).
46. Zhou, J. *et al*. Chromosome-level genome assembly reveals significant gene expansion in the Toll and IMD signaling pathways of *Dendrolimus kikuchii*. *Front. Genet.* **12**, 728418, https://doi.org/10.3389/fgene.2021.728418 (2021).
47. Kawamoto, M. *et al*. High-quality genome assembly of the silkworm, *Bombyx mori*. *Insect Biochem. Mol. Biol.* **107**, 53–62, https://doi.org/10.1016/j.ibmb.2019.02.002 (2019).
48. Wan, W. *et al*. Genome-wide survey of open chromatin regions in two swallowtail butterflies *Papilio machaon* and *P. bianor*. *Arch. Insect Biochem. Physiol.* **111**, e21952, https://doi.org/10.1002/arch.21952 (2022).
49. Boyes, D. *et al*. The genome sequence of the Diamondback Moth, *Plutella xylostella* (Linnaeus, 1758). *Wellcome Open Res.* **8**, 404, https://doi.org/10.12688/wellcomeopenres.20006.1 (2023).
50. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 1–14, https://doi.org/10.1186/s13059-019-1832-y (2019).
51. Katoh, K., Asimenos, G. & Toh, H. Multiple alignment of DNA sequences with MAFFT. *Methods Mol. Biol.* **537**, 39–64, https://doi.org/10.1007/978-1-59745-251-9_3 (2009).
52. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56**, 564–577, https://doi.org/10.1080/10635150701472164 (2007).
53. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274, https://doi.org/10.1093/molbev/msu300 (2015).
54. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589, https://doi.org/10.1038/nmeth.4285 (2017).
55. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556, https://doi.org/10.1093/bioinformatics/13.5.555 (1997).
56. Han, M. V., Thomas, G. W., Lugo-Martinez, J. & Hahn, M. W. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol. Biol. Evol.* **30**, 1987–1997, https://doi.org/10.1093/molbev/mst100 (2013).
57. *NCBI National Genomics Data Center(NGDC) database* https://identifiers.org/ncbi/bioproject:PRJNA1289346 (2025).
58. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR34562996 (2025).
59. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR34567164 (2025).
60. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR34577075 (2025).
61. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR34587760 (2025).
62. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBPQDY020000000 (2026).
63. Wang, Z., Kong, D. & Qu, L. Genome assembly and annotation of the teak defoliator *Hyblaea puera* Cramer. figshare. *Dataset.* https://doi.org/10.6084/m9.figshare.31175569.v1 (2026).
64. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**(18), 3094–3100, https://doi.org/10.1093/bioinformatics/bty191 (2018).
65. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212, https://doi.org/10.1093/bioinformatics/btv351 (2015).

Acknowledgements
This study was supported by the National Key R&D Program of China (2023YFC2604800), and the Fundamental Research Funds for the Central Non-profit Research Institution of Chinese Academy of Forestry (CAFYBB2024QF034).

Author contributions
L.Q. conceived the research idea; Z.W., D.K., Y.L., T.Z., Q.W., J.Z. and Y.Z. participated in specimen collection and sample preparation; Z.W. and D.K. performed the data analysis; Z.W. and L.Q. wrote the manuscript. All authors read, revised, and approved the final version of the manuscript.

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/s41597-026-07022-8.

Correspondence and requests for materials should be addressed to L.Q.