scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of the casuarina moth, *Lymantria xylina* Swinhoe (1903)

Siyi Liu, Hui Jiang, Tao Ni, Kai Wang, Xia Hu, Songqing Wu, Feiping Zhang & Rong Wang

The casuarina moth, *Lymantria xylina*, is a serious pest threatening subtropical regions through severe defoliation and strong invasive potential. Despite its economic impact and high invasion risk, a high-quality reference genome remains lacking. To bridge this knowledge gap, we generated a chromosome-level genome assembly for *L. xylina* combining Illumina short-reads, Oxford Nanopore long-reads, and high-throughput chromatin conformation capture (Hi-C) scaffolding data. Following long-reads based assembly and Hi-C scaffolding, the final genome assembly totals 977.74 Mb, with 930.50 Mb (95.17%) of sequences anchored onto 31 pseudo-chromosomes, achieving a scaffold N50 of 34.15 Mb. The genome assembly, featuring fully assembled telomeres on all 31 pseudo-chromosomes, demonstrates 94.5% Benchmarking Universal Single-Copy Orthologs (BUSCO) completeness and high accuracy with consensus quality value of 31.72. Repetitive elements constitute 77.18% of the genome, and 18,484 protein-coding genes were predicted, with 95.21% functionally annotated. This high-quality genome assembly provides a critical foundation for elucidating interaction mechanisms with host plants and natural enemies (nucleopolyhedrovirus, *Beauveria bassiana*), for developing enhanced pest management and control strategies.

Background & Summary

Lymantria xylina Swinhoe (1903), commonly known as the casuarina moth, is a significant forestry pest native to Japan, India, and the eastern coast of mainland China including Taiwan[1]. This polyphagous lepidopteran is univoltine, exhibiting distinct phenology. The adults emerge in late May to early June (Japan) or late May–late June (China), with males preceding females by several days[2,3]. After mating, females deposit light brown, setae-covered egg masses (200–1500 eggs; max. 2435 eggs) 2–4 m aboveground without host preference[1,4,5]. Eggs undergo diapause as pharate first instars[6], mainly hatching in April after 8–9 months of dormancy. Larvae (45–47 mm) disperse via silk ballooning and display extreme polyphagy, expanding their host range with age[2,5]. The larval stage (5–7 instars) lasts 1.5–2 months, culminating in pupation on tree limbs/leaves (5–14 days) before adult eclosion[7].

This polyphagous lepidopteran is notorious for causing severe defoliation outbreaks and infests a wide range of economically important hosts, fruit trees (*Dimocarpus longan*, *Litchi chinensis*, *Mangifera indica, etc*), and other hardwoods[7]. Its high invasiveness risk, facilitated by natural dispersal (larval ballooning, adult flight) and accidental human-mediated transport (egg masses on shipping containers)[8–10], poses a significant threat to subtropical regions. Despite its ecological and economic impact, existing research has primarily focused on its ecology, phenology, and control methods, while a high-quality chromosome-level genome remains unavailable. This gap hinders deeper molecular studies into its adaptation, invasiveness, and interactions with host plants and natural enemies like nucleopolyhedrovirus[11] and *Beauveria bassiana*[12]. To bridge this knowledge gap, we generated a chromosome-level genome assembly for *L. xylina* combining Illumina short-reads, Oxford Nanopore long-reads, and Hi-C scaffolding technique.

The generated sequencing data volumes were summarized in Table 1. K-mer analysis of 51.33 Gb Illumina data identified 43.98 billion 17-mers with main peak depth of 46, providing initial insights into the genomic characteristics of *L. xylina* (Table 2). The genome size of *L. xylina* was estimated to 869.46 Mb, with GC content of 39.34%, repeat content of 60.25% and heterozygosity of 0.65%. Subsequent *de novo* assembly of Nanopore long-reads yielded a 977.73 Mb draft genome with contig N50 of 11.05 Mb and maximum contig

State Key Laboratory of Agricultural and Forestry Biosecurity, College of Forestry, Fujian Agriculture and Forestry University, Fuzhou, 350002, China. ✉e-mail: fpzhang1@163.com; wangrongbl0501@126.com

Data Type	Platform	Reads length (bp)	Clean reads	Total clean reads bases (Gb)
Illumina paired-end short reads	Illumina NovaSeq 6000	2 × 150	342,751,748	51.33
Nanopore long reads	Nanopore PromethION	29,214 (mean length)	3,843,216	112.28
Hi-C reads	Illumina NovaSeq 6000	2 × 150	431,775,903	129.28
RNA-seq reads	Illumina NovaSeq 6000	2 × 150	60,389,720	4.52

Table 1. Summary of high throughput sequencing data generated in this study.

Clean data (Gb)	Depth (X)	kmer size	Number of kmer	Genome size (Mb)	GC content (%)	Repeat (%)	Heterozygous ratio (%)
51.33	46	17	43,980,308,637	869.46	39.34	60.25	0.65

Table 2. K-mer analysis of the *L. xylina* genome.

Characteristics	Value
Assembled genome features	
Contig level	
Assembled genome size (Mb)	977.73
GC content	38.63
Number of contigs	1,189
Longest contigs (Mb)	38.94
N50 of long reads assembly (Mb)	11.05
N90 of long reads assembly (Mb)	2,49
Chromosome level	
Total assembled genome size (Mb)	977.74
Pseudo-chromosomes number	31
Sequence anchored to pseudo-chromosomes (Mb)	930.50 (95.17%)
Number of scaffolds	1,071
N50 of contigs (Mb)	10.16
N50 of scaffolds (Mb)	34.15
N90 of scaffolds (Mb)	16.36
Genome completeness and quality	
BUSCO (lepidoptera_odb12)	94.5% [S:93.60%, D:0.9%, F:1.4%, M:4.1%]
Merqury (QV value)	31.72 [Single base error rate: 0.00067367]
Genome annotation	
Number of predicted protein-coding genes	18,484
Average gene length (bp)	13,174
Percentage of functional annotation genes (%)	95.21
Percentage of repeat sequences (%)	77.18%

Table 3. Summary of *L. xylina* genome assembly and annotation.

length of 38.94 Mb (Table 3). A cumulative of 930.50 Mb (95.17%) of sequences were successfully anchored and oriented onto 31 pseudo-chromosomes by Hi-C valid interaction links (Table 4). Finally, a high-quality chromosome-level genome was obtained, showing scaffold N50 of 34.15 Mb, with a total size of 977.74 Mb (Table 3). Quality assessment demonstrated high completeness (94.50% BUSCO score using lepidoptera_odb12) and high accuracy (Merqury QV 31.72). Through integrated *de novo* and homology-based approaches, repetitive sequences were found to constitute 77.18% of the *L. xylina* genome assembly (Table 3). A total of 18,484 protein-coding genes were predicted, 95.21% of which were successfully functionally annotated (Table 3).

Overall, this high-quality chromosome-level genome, featuring a scaffold N50 of 34.15 Mb and 18,484 functionally annotated protein-coding genes, provides a critical and specific foundation for mechanistic research. Specifically, the comprehensive gene set offers a powerful resource to elucidate the genetic basis of key biological traits: by identifying genes associated with polyphagy (e.g., detoxification enzymes[13]), investigating the regulation of its obligate egg diapause[14], and determining the molecular pathways underlying its high invasiveness[15]. This genomic resource will directly inform the design of targeted control strategies, including deeper molecular studies into its interactions with host plants and potential biocontrol agents[16].

Data	Value
Statistics of Hi-C data	
Clean data (bp)	129,281,643,972
Total read pairs	431,775,903
GC (%)	44.7
>Q30 (%)	92.55
Mapping stats of Hi-C data	
Total mapped reads	749,868,077
Unique mapped read pairs	249,556,026
Percentage of unique mapped read pairs (%)	57.80
Statistics of valid Hi-C data	
Valid interaction pairs	82,664,649
Percentage of valid interaction read pairs (%)	33.12
Dangling End Pairs	149,447,615
Re-ligation Pairs	2,319,123
Self-cycle Pairs	472,270
Dumped Pairs	14,652,369

Table 4. Statistics of Hi-C data and valid data.

Methods

Sample collection and preparation. For whole-genome sequencing, the egg masses of *L. xylina* were obtained from a casuarina plantation in Pingtan Comprehensive Experimental Zone, Fuzhou City, Fujian Province (25°27′02.9″N 119°47′58.3″E) and were maintained at 4 °C before hatching. Hatched larvae were fed with an artificial diet at 25 ± 1 °C, 14:10 (L:D) photoperiod and 65 ± 5% relative humidity. The larvae, pupae, and adult moth were collected separately. The samples were frozen in liquid nitrogen and then stored at −80 °C.

For transcriptome sequencing, field-collected *L. xylina* eggs from the same plots were staged, early-diapause (ED group) at 7 days post-oviposition (dpo) with milky-white chorions, mid-diapause (MD group) at 30 dpo exhibiting embryonic eye-spot formation. For diapause termination (DT group), MD eggs received 4 °C/60-day treatment. DT eggs were transferred to 25 °C/75% relative humidity developed to pre-hatching (DTH group) within 72 ± 3 hours, evidenced by larval mandibular sclerotization. Each group (ED/MD/DT/DTH) contained three replicates, all twelve biological samples underwent parallel processing for transcriptomic and metabolomic profiling. For metabolic analysis, samples from the four groups (DT, DTH, ED, and MD), with each group comprising five biological replicates, were thawed on ice. Take 50 ± 2 mg of one sample and add cold steel balls to the mixture and homogenate at 30 Hz for 3 min. Add 1 mL 70% methanol with internal standard extract to the homogenized centrifuge tube, Whirl the mixture for 5 min, and then centrifuge it with 12,000 rpm at 4 °C for 10 min. After centrifugation, draw 400 uL of supernatant into the corresponding EP tube and store in −20 °C refrigerator overnight, then centrifuge at 12000 r/min at 4 °C for 3 min, and take 2000 uL of supernatant in the liner of the corresponding injection bottle for on-board analysis.

Whole genome sequencing. For short reads sequencing, two DNA-seq libraries with 350-bp inserts were constructed and subjected to paired-end sequencing (2 × 150 bp) on the Illumina NovaSeq 6000 platform, yielding approximately 51.33 Gb of data (Table 1). To construct Nanopore sequencing libraries, 8 μg of genomic DNA (gDNA) was size-selected using the Blue Pippin system (Sage Science, USA). Subsequently, the LSK-109 ligation sequencing kit (SQK-LSK109, Oxford Nanopore Technologies) was used for adapter ligation. The prepared libraries were then sequenced on the PromethION platform (Oxford Nanopore Technologies, UK). This generated 112.28 Gb of long-read sequence data, comprising 3,843,216 reads with a mean length of 29.21 kb (Table 1). To achieve a chromosome-level genome assembly, two Hi-C (high-throughput chromatin conformation capture) libraries were constructed and sequenced on the Illumina NovaSeq 6000 platform (Illumina, USA). Raw sequencing data underwent quality control using fastp (v0.21.0)[17] to remove adapter-containing and low-quality reads, generating 129.28 Gb of high-quality data (Table 1). Furthermore, one RNA-seq library was constructed and subjected to 150-bp paired-end sequencing on the Illumina NovaSeq 6000 platform. Post-quality control, 4.52 Gb of high-quality data were generated to support protein-coding gene annotation (Table 1).

Estimation of genomic characteristics using k-mer analysis. The raw short reads underwent quality filtering using fastp (v0.21.0)[17]. Subsequently, the clean reads were employed for genome size estimation. Firstly, a total of 43,980,308,637 distinct 17-mers were identified from 51.33 Gb of high-quality short-read data using Jellyfish (v2.3.0)[18] (Table 2). The k-mer frequency distribution was subsequently modeled with GenomeScope2 (v2.0)[19] under a diploid model (-p 2) using k-mer size 17 and 3 refinement rounds (–num_rounds 3) to estimate genome properties, with main depth peak at 46 (Fig. 1 and Table 2). This revealed an estimated genome size of 869.46 Mb characterized by high repeat content (60.25%) and moderate heterozygosity (0.65%).

Chromosome-level genome assembly. After performing error correction on Nanopore third-generation sequencing data using Canu (v1.6)[20], high-accuracy long reads were obtained. *De novo* assembly was then conducted using this error-corrected data. The assembly was then refined through two rounds of polishing with

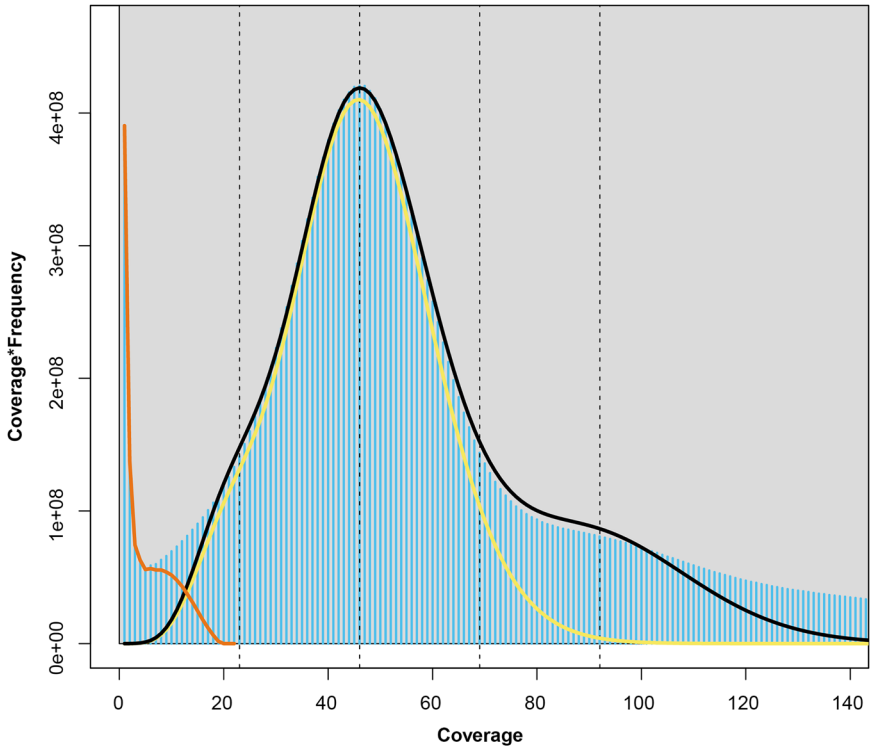

Fig. 1 Genome survey analysis of *L. xylina* using k-mer frequency distribution (k = 17). The x-axis indicates k-mer depth and the y-axis indicates the frequency of k-mers at each corresponding depth.

third-generation sequencing data using Racon (v1.5.0)[21], followed by three additional rounds of polishing with second-generation sequencing data using Pilon (v1.24)[22]. The final assembled genome had a total length of 977.73 Mb, with a contig N50 of 11.05 Mb and the longest contig spanning 38.94 Mb. Leveraging genome-wide spatial interaction information from Hi-C data, we refined, clustered, and oriented the contig-level genome assembly to achieve chromosome-level resolution. A total of 129.28 Gb of clean Hi-C data (431.78 million paired reads) was filtered and evaluated using HiC-Pro (v3.1.0)[23], with 57.80% of reads uniquely mapping to the contig-level genome (Table 4). The construction of 31 pseudo-chromosomes using 33.12% of valid Hi-C contacts was executed through Lachesis[24]. In total, 930.50 Mb (95.17%) of sequences were anchored onto 31 pseudo-chromosomes, yielding a final genome size of 977.74 Mb with a scaffold N50 of 34.15 Mb (Table 3). The longest pseudo-chromosome measured 45.21 Mb, while the shortest was 11.90 Mb (Table 5). Telomere identification performed by quarTeT (v1.1.7)[25] confirmed telomeric sequences at both ends of all 31 pseudo-chromosomes (Table 5). Intra- and inter-chromosomal interaction heatmaps were visualized using a custom script at 500 Kb resolution (Fig. 2).

Repetitive sequences annotation. Due to the species-specific distribution of repetitive sequences, the construction of a species-specific repeat database is essential for accurate prediction. Therefore, *L. xylina*-specific repeat library was generated through *de novo* prediction using LTR_FINDER (v1.0.7)[26] and RepeatScout (v1.0.6)[27] based on structural features. The resulting sequences were classified with PASTEClassifier (v2.0)[28] and subsequently merged with the Repbase TE library (version 15.02)[29] to produce a consolidated repeat library. Repeat annotation was then performed against this integrated database using RepeatMasker (v4.1.2)[30]. A total of 754.59 Mb repetitive sequences were identified, accounting for 77.18% of the *L. xylina* genome. Among these repetitive elements, long terminal repeats (LTRs) and DNA transposons were identified as the predominant types (Table 6).

Protein coding genes prediction and functional annotation. Prediction of protein-coding genes in the *L. xylina* genome was conducted through an integrated approach combining homology-based, *ab initio*-based, and transcriptome-based methodologies. Homology predictions utilized protein sequences from *Bombyx mori*, *Apis mellifera*, *Plutella xylostella*, and *Drosophila melanogaster* downloaded from RefSeq database, with genomic alignments constructed using GeMoMa (v1.7)[31]. *Ab initio* predictions employed multiple tools including Augustus (v3.1.0)[32], GlimmerHMM (v3.0.4)[33], GeneID (v1.4)[34], Genscan[35], and SNAP (2006-07)[36]. For transcriptome-assisted prediction, clean RNA-seq reads were aligned to the genome using HiSat2 (v2.0.4)[37] and assembled into transcripts via StringTie (v1.2.3)[38]. These transcripts were directly utilized for gene prediction with GeneMarkS-T (v5.1)[39] and TransDecoder (v2.0) (https://github.com/TransDecoder/io), while *de novo* transcriptome assemblies were annotated using PASA (v2.4.1)[40]. All evidence sources were integrated

Chromosome	Length (bp)	Telomere status	Left	Direction of left	Right	Direction of right
LG01	45,208,564	both	113	+	88	+
LG02	35,865,157	both	107	–	83	–
LG03	41,941,757	both	92	–	98	–
LG04	39,944,401	both	84	+	68	–
LG05	39,595,413	both	86	+	79	–
LG06	39,782,319	both	95	+	76	+
LG07	38,938,878	both	103	+	109	–
LG08	38,104,242	both	102	+	117	–
LG09	37,632,118	both	83	+	109	–
LG10	37,520,216	both	79	+	101	–
LG11	36,913,288	both	98	–	104	–
LG12	33,718,594	both	89	+	96	–
LG13	35,524,488	both	106	-	91	+
LG14	33,915,424	both	93	+	86	–
LG15	34,150,346	both	121	–	114	–
LG16	29,341,851	both	112	–	96	+
LG17	30,995,287	both	69	+	119	–
LG18	30,384,833	both	81	–	73	–
LG19	29,965,776	both	100	+	80	+
LG20	27,998,077	both	83	+	81	–
LG21	28,050,024	both	113	+	70	+
LG22	23,676,047	both	86	+	76	–
LG23	23,816,463	both	100	-	92	+
LG24	22,495,500	both	69	+	158	+
LG25	20,637,702	both	140	–	83	+
LG26	20,367,808	both	126	+	87	–
LG27	16,360,031	both	129	+	104	+
LG28	16,432,268	both	94	–	104	–
LG29	11,896,210	both	107	–	100	+
LG30	16,152,624	both	128	+	78	–
LG31	13,189,435	both	105	–	111	+

Table 5. Telomere repeat monomer distribution in *L. xylina*.

by EvidenceModeler (v1.1.1)[41] to generate a non-redundant gene set, with PASA further refining terminal exon structures to optimize gene models. This pipeline identified 18,484 protein-coding genes (Table 7) with an average gene length of 13.17 kb (Table 3). The distribution of GC content, repetitive sequences and protein-coding genes on each chromosome of *L. xylina* was visualized using a circos plot (Fig. 3).

The protein sequences were mapped to the NR database (https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/) using BLAST (e-values $\leq$ 1e−5)[42], and the closest hit was selected as the annotation result. In addition, KOG[43] and TrEMBL[44] databases were used for a comprehensive functional annotation of genes. The process of alignment was based on Diamond (v2.0.14)[45]. Blast2GO was used to determine the functions and pathways based on the GO and KEGG databases[46]. Among all protein-coding genes, 95.21% of them were successfully annotated (Table 8).

Non-coding RNA annotation. Transfer RNAs (tRNAs) were predicted using tRNAscan-SE (v1.3.1)[47], while identification of ribosomal RNAs (rRNAs) and microRNA (miRNA) were generated through BLASTN alignments against the Rfam database (v14.5)[48]. Comprehensive analysis revealed the *L. xylina* genome contained 680 tRNA genes, 118 rRNA genes, and 38 miRNAs (Table 9).

RNA libraries construction, sequencing, and data analysis. The 12 samples from four experimental groups (ED, MD, DT, and DTH, with three biological replicates per group) underwent rigorous RNA quality assessment. Sequencing libraries were constructed from 1 μg total RNA per sample using the NEBNext® Ultra™ RNA Library Prep Kit for Illumina®. Key steps included poly-A mRNA selection, fragmentation, double-stranded cDNA synthesis, adapter ligation, size selection (250-300 bp), and PCR amplification. Final libraries were quality-checked, and sequenced on a HiSeq platform to generate 150 bp paired-end reads (Table 10). For data analysis, raw reads were quality filtered using fastp (v0.19.3) to remove adapter sequences, reads with >10% N bases, and reads where >50% of bases had Q $\leq$ 20, resulting in clean reads. These clean reads were aligned to the designated reference genome using Hisat2[49]. Novel transcripts were predicted using StringTie (v1.3.4 d)[38].

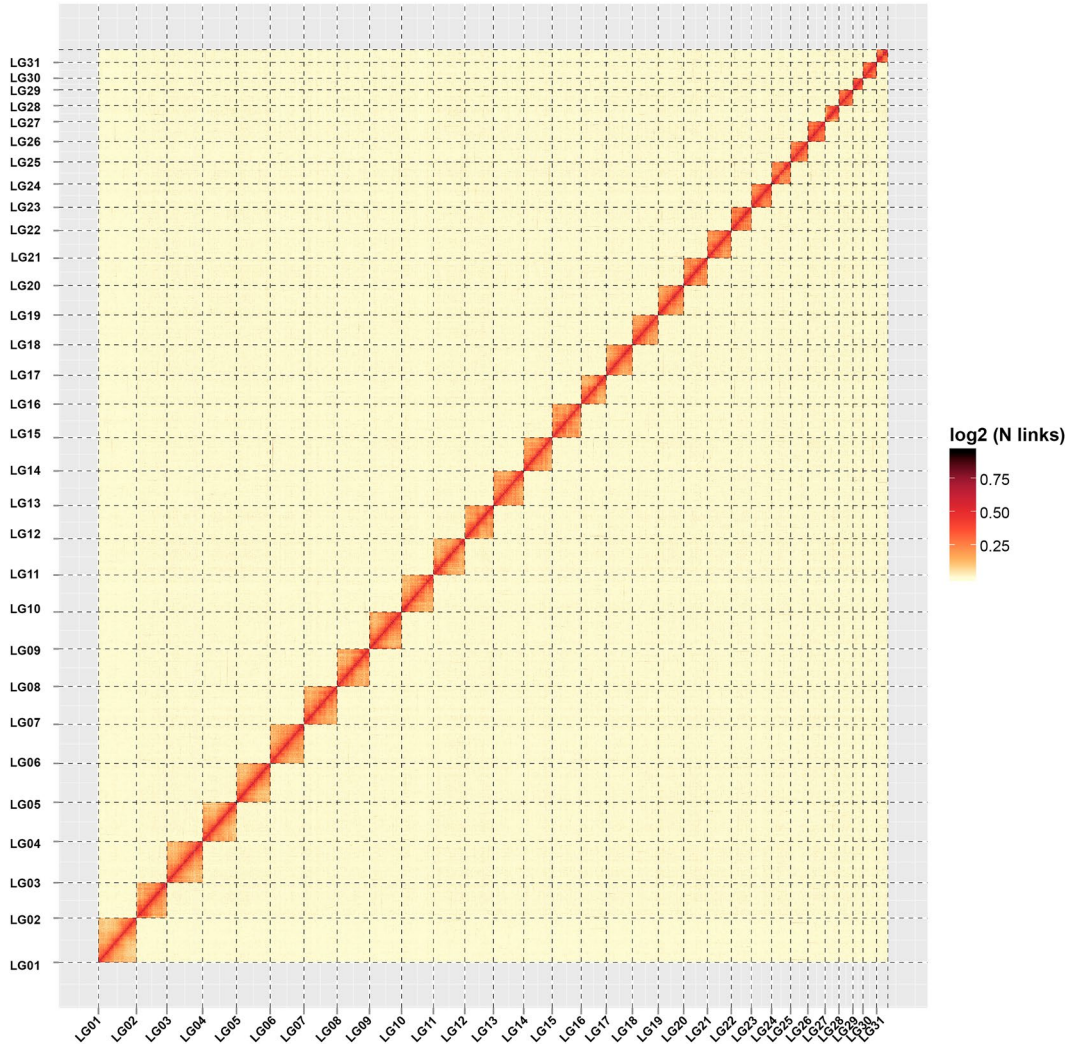

Fig. 2 Heatmap showing Hi-C interaction frequencies across 31 pseudo-chromosomes of *L. xylina* at 500-kb resolution. The binary logarithm of Hi-C link counts is represented using a color gradient from light yellow to dark red, reflecting interaction frequencies from lowest to highest. LG01 to LG31 designate the 31 pseudo-chromosomes.

Gene expression levels were quantified as FPKM using featureCounts (v1.6.2) to calculate gene alignments[50]. Significantly differential expression analysis between groups was performed using DESeq2 (v1.22.1)[51], with adjusted *P*-value ≤ 0.05 and absolute $\log_2$(fold change) ≥ 1. The pairwise comparison between ED and DTH groups exhibited the highest number of differentially expressed genes (DEGs) (n = 8,910), while DT versus DTH showed the fewest DEGs (n = 780) (Table 11).

Integrated metabolite identification strategy using UPLC-MS/MS

Metabolite identification was conducted using ultra-performance liquid chromatography coupled to tandem mass spectrometry (UPLC-MS/MS). Chromatographic separation was achieved on a Waters ACQUITY UPLC HSS T3 C18 column with gradient elution (5%–90% acetonitrile containing 0.1% formic acid) over a 14-minute runtime. Complementary mass spectrometry platforms were employed: untargeted screening utilized a TripleTOF 6600 system performing full scans and information-dependent acquisition (IDA) of MS/MS spectra for high-intensity precursors, while targeted verification used a QTRAP 5500 system operating in multiple reaction monitoring (MRM) mode. Both MS systems operated with electrospray ionization (ESI) under standardized source conditions, including polarity switching (±5500 V), 500 °C source temperature, and identical gas pressure settings for ion source and curtain gases.

Identification of differential expressed metabolites (DEMs)

Significantly regulated metabolites between groups were determined by variable importance in projection (VIP) ≥ 1 and absolute $\log_2$FC (fold change) ≥ 1. VIP values were extracted from OPLS-DA result, which also contain score plots and permutation plots, was generated using R package MetaboAnalystR[52]. The data underwent $\log_2$ transformation and mean centering before OPLS-DA analysis. In order to avoid overfitting,

Type	Number	Total length (bp)	Percentage of genome (%)
ClassI:Retroelement	3,461,450	620,090,305	63.42
LTR-retroelement			
LTR/Copia	6,058	2,810,309	0.29
LTR/Gypsy	55,744	25,540,610	2.61
LTR/Others	52,935	25,512,983	2.61
Non LTR-Retroelement			
LINE	2,096,601	433,739,014	44.36
SINE	1,156	209,043	0.02
DIRS	7,378	2,692,713	0.28
TRIM	1,301	898,023	0.09
PLE|LARD	1,240,137	261,980,023	26.79
Other Retroelement	140	36,550	0
ClassII:DNA transposon	1,040,352	253,108,408	25.89
TIR	944,056	230,123,797	23.54
Helitron	63,671	17,352,864	1.77
Maverick	16,780	3,766,376	0.39
MITE	4,084	3,555,469	0.36
Others			
Unknown	11,701	3,149,031	0.32
Unknown	259,591	58,033,052	5.94
Total	4,512,319	754,587,205	77.18

Table 6. Annotation of repetitive elements in the *L. xylina* genome.

Strategy	Software	Species	Protein coding gene number
Ab initio			
	Genscan	—	13,710
	Augustus	—	16,871
	GlimmerHMM	—	80,569
	GeneID	—	13,657
	SNAP	—	20,123
Homology-based			
	GeMoMa	*Bombyx mori*	14,535
		Apis mellifera	7,976
		Plutella xylostella	14,579
		Drosophila melanogaster	7,748
RNA-seq	TransDecoder	—	34,992
	GeneMarkS-T	—	23,219
	PASA	—	10,218
Integration	EVM	—	18,484

Table 7. Prediction of protein-coding genes in the *L. xylina* genome.

a permutation test (200 permutations) was performed. Finally, pairwise comparisons revealed that MD and DTH groups exhibited the highest number of DEMs (n = 553), while DT vs DTH showed the fewest (n = 31) (Table 12).

Data Records

Whole-genome sequencing reads have been deposited in Genome Sequence Archive (GSA)[53] at National Genomics Data Center (NGDC)[54], China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences (GSA: CRA027397)[55] that are publicly accessible at https://ngdc.cncb.ac.cn/gsa. Additionally, raw high-throughput sequencing data for *L. xylina* have been deposited in the NCBI Sequence Read Archive with accession number SRP655858[56]. The chromosome-level genome assembly of *L. xylina* has been deposited in the Genome Warehouse at NGDC under accession number GWHGEMT00000000.1, that is publicly accessible at https://ngdc.cncb.ac.cn/gwh[57]. The assembly is also available in GenBank under accession number JBPSJU000000000[58]. Additionally, the genome annotation files and detailed information on differentially expressed genes and metabolites across different groups are available from Figshare[59]. All untargeted

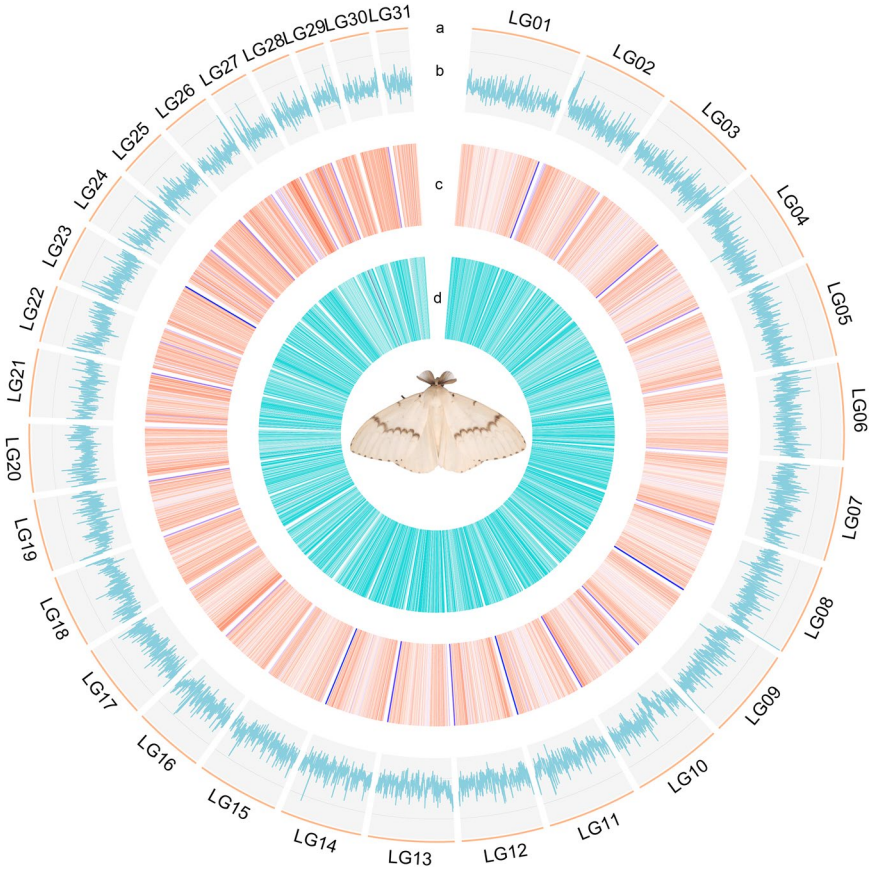

Fig. 3 Genomic characteristics of the 31 pseudo-chromosomes of *L. xylina*. Tracks a-d depict chromosome ideograms, GC content, gene density, and repetitive sequence density, all calculated within 100-kb windows.

Type	Number of genes	Percentage of genes (%)
Total annotated genes	17,598	95.21
GO	8,827	47.75
KEGG	6,709	36.30
KOG	11,162	60.39
Pfam	10,651	55.58
InterPro	11,967	62.44
Swiss-Prot	7,824	40.82
TrEMBL	17,341	93.82
NR	17,425	94.27

Table 8. Results of functional annotation of predicted protein-coding genes.

Type	Number	Family
rRNA	118	4
tRNA	680	25
miRNA	38	30

Table 9. Prediction of non-coding RNAs in the *L. xylina* genome.

metabolomic data used in this publication have been deposited to the EMBL-EBl MetaboLights database with the identifier MTBLS13575[60]. The complete data set can be accessed at https://www.ebi.ac.uk/metabolights/MTBLS13575.

Group	Sample	Raw Reads	Clean Reads	Clean Base (Gb)	Error Rate (%)	Q20 (%)	Q30 (%)	GC (%)
DT	DT1	50,218,668	47,867,460	7.18	0.03	97.91	93.91	42.67
DT	DT2	48,312,868	46,078,692	6.91	0.02	98.08	94.29	44.16
DT	DT3	48,330,628	46,345,736	6.95	0.03	97.97	94.06	42.25
DTH	DTH1	47,534,264	45,665,642	6.85	0.02	98.15	94.44	42.35
DTH	DTH2	50,265,462	48,362,764	7.25	0.02	98.07	94.19	42.13
DTH	DTH3	45,759,468	43,925,566	6.59	0.02	98.13	94.39	42.39
ED	ED1	46,976,940	45,080,012	6.76	0.02	98.17	94.47	41.43
ED	ED2	43,476,068	41,445,670	6.22	0.03	98.05	94.15	41.29
ED	ED3	49,257,392	47,130,964	7.07	0.02	98.07	94.24	43.17
MD	MD1	47,837,664	45,705,742	6.86	0.02	98.2	94.56	42.31
MD	MD2	46,633,248	44,239,076	6.64	0.02	98.1	94.3	41.11
MD	MD3	44,674,128	42,881,464	6.43	0.02	98.15	94.33	40.83

Table 10. Summary of RNA-seq data for *L. xylina* eggs across different groups.

Group1	Group2	Number of DEGs	Up-regulated	Down-regulated
DT	DTH	780	467	313
DT	MD	3,357	1,975	1,382
DT	ED	8,045	4,262	3,783
ED	DTH	8,910	4,723	4,187
ED	MD	3,546	2,084	1,462
MD	DTH	4,651	2,730	1,921

Table 11. Summary of DEGs in pairwise comparisons among *L. xylina* experimental groups.

Group1	Group2	Number of DEMs	Up-regulated	Down-regulated
DT	DTH	31	16	15
DT	MD	534	299	235
DT	ED	401	355	46
ED	DTH	427	369	58
ED	MD	511	326	185
MD	DTH	553	310	243

Table 12. Summary of DEMs in pairwise comparisons among *L. xylina* experimental groups.

Technical Validation

Various different strategies were used to evaluate the completeness and accuracy of *L. xylina* genome. Firstly, genome completeness was evaluated based on BUSCO (v5.8.1)[61] with the lepidoptera_odb12 database, which achieved an assembly completeness of 94.5% (S:93.60%, D:0.9%, F:1.4%, M:4.1%) (Table 3). Secondly, an assessment of the consistency quality value (QV) for *L. xylina* was conducted using Merqury[62]. The results indicated that the QV for our chromosomal version of the genome was 31.72, suggesting a genome-wide error rate of only 0.00067367 (Table 3). Thirdly, the Hi-C interaction heatmap revealed pronounced intra-chromosomal interactions with minimal inter-chromosomal signal (Fig. 2). Simultaneously, telomeric sequences were identified at both ends of all chromosomes (Table 5). These evidences collectively confirmed a haploid chromosome number of 31 in *L. xylina*. Lastly, Finally, we employed SyRI (v1.7.1)[63] to assess synteny between *L. xylina* and its congeneric species Asian spongy moth (*L. dispar asiatica*, GCA_032191425.1)[64] (Fig. 4). The analysis revealed that a substantial portion of the genomes in these two related species (726 Mb in *L. xylina* and 712 Mb in *L. dispar asiatica*) was structurally conserved (syntenic blocks), with conserved gene order and orientation (Table 13). All these findings suggested this *L. xylina* assembled genome was of high quality, characterized by its high accuracy, completeness, contiguity, and substantial coverage.

Data availability

The raw sequencing reads have been deposited in GSA (CRA027397, https://ngdc.cncb.ac.cn/gsa). Additionally, raw high-throughput sequencing data for *L. xylina* have been deposited in the NCBI Sequence Read Archive with accession number SRP655858. The genome assembly of *L. xylina* is available in GWH (GWHGEMT00000000.1, https://ngdc.cncb.ac.cn/gwh) and GenBank (JBPSJU000000000). Genome annotation and related data are available at Figshare (https://doi.org/10.6084/m9.figshare.29497898). All untargeted metabolomic data used in this publication have been deposited to the EMBL-EBl MetaboLights database with the identifier MTBLS13575

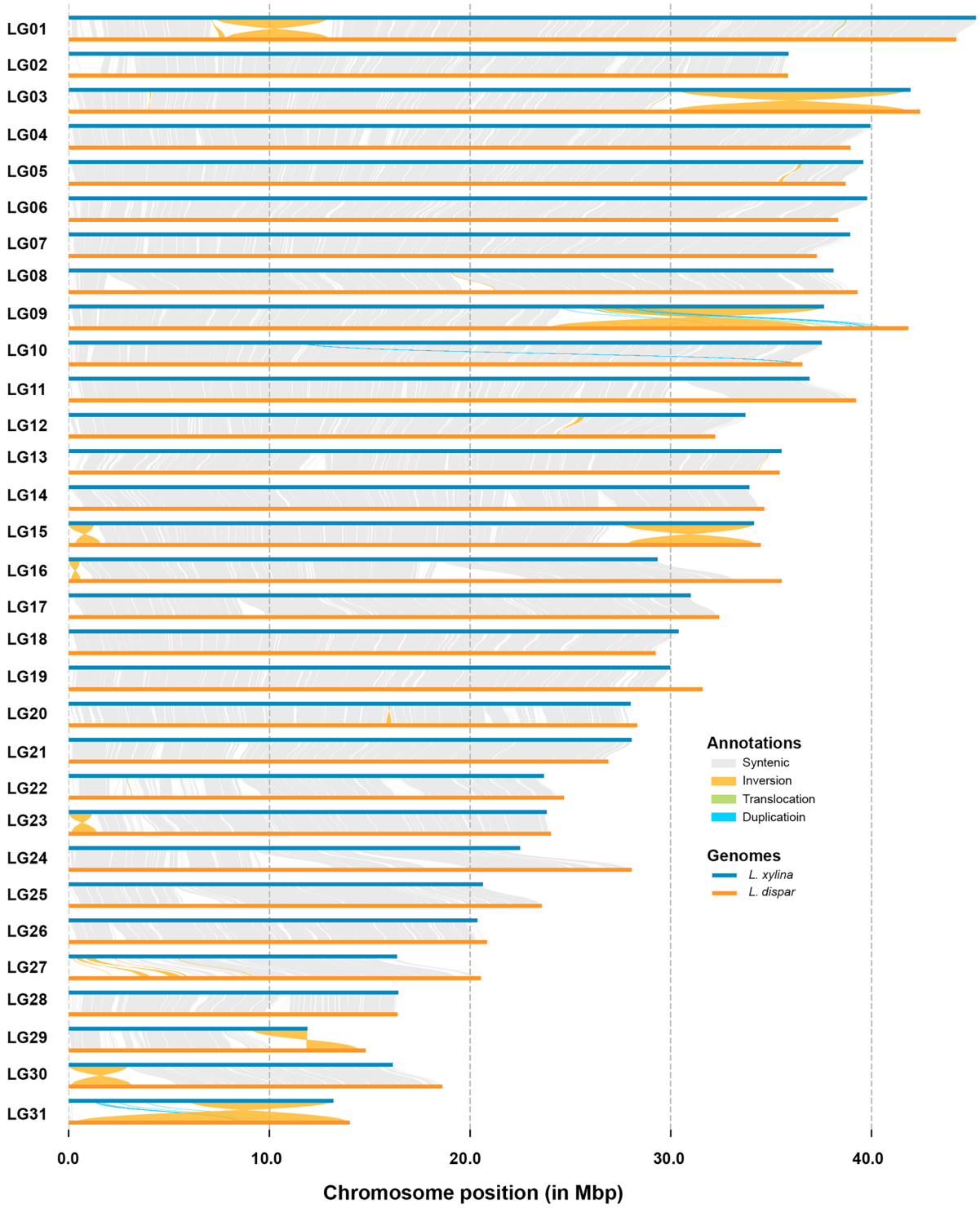

Fig. 4 Comparative genomic structural variations between *L. xylina* and *L. dispar*. Pseudo-chromosomes (LG01–LG31) are shown with chromosomal positions (in megabase pairs, Mbp). Annotations highlight synthetic regions, inversions, translocations, and duplications identified across the two genomes. Scale bar indicates genomic position (0–40 Mbp).

(https://www.ebi.ac.uk/metabolights/MTBLS13575). Additionally, the processed metabolomics dataset, including metabolite identification tables and quality control metrics, has been deposited in Figshare at https://doi.org/10.6084/m9.figshare.30909260.

Variation_type	Count	Length in *L. xylina*	Length in *L. dispar*
Syntenic regions	3,475	726,081,990	712,468,394
Inversions	300	52,229,851	60,733,290
Translocations	3,207	6,185,050	6,187,680
Duplications (*L. xylina*)	713	1,967,002	—
Duplications (*L. dispar*)	2,722	—	6,896,100
Not aligned (*L. xylina*)	7,650	144,239,643	—
Not aligned (*L. dispar*)	9,523	—	172,575,545

Table 13. Whole-genome synteny statistics between *L. xylina* and *L. dispar*.

Code availability

No custom code was used for this study. All data analyses were performed using published bioinformatics software with specified parameter settings, as stated in the Methods section and Figshare[65].

Received: 25 September 2025; Accepted: 26 January 2026;
Published online: 04 February 2026

References

1. Wallner, W. E. & McManus, K. A. Proceedings, Lymantriidae: A comparison of features of New and Old World tussock moths. *Gen. Tech. Rep. NE-123. Broomall, PA: US Department of Agriculture, Forest Service, Northeastern Forest Experiment Station. 554 p.* **123**, 1-554 (1989).
2. Shen, T.-C., Tseng, C.-M., Guan, L.-C. & Hwang, S.-Y. Performance of *Lymantria xylina* (Lepidoptera: Lymantriidae) on artificial and host plant diets. *J Econ Entomol* **99**, 714–721 (2006).
3. Pogue, M. A review of selected species of *Lymantria* Hübner (1819)(Lepidoptera: Noctuidae: Lymantriinae) from subtropical and temperate regions of Asia, including the descriptions of three new species, some potentially invasive to North America. (U.S. Department of Agriculture, Forest Health Technology Enterprise Team, 2007).
4. Hwang, F. Investigation of the larval food plant range of the casuarina moth, *Lymantria xylina*. (Doctoral dissertation, MS thesis, National Chung Hsing University. Taichung, Taiwan, 2005).
5. Hwang, S.-Y., Hwang, F.-C. & Shen, T.-C. Shifts in developmental diet breadth of *Lymantria xylina* (Lepidoptera: Lymantriidae). *J Econ Entomol* **100**, 1166–1172 (2007).
6. Hwang, S.-Y., Jeng, C.-C., Shen, T.-C., Shae, Y.-S. & Liu, C.-S. Diapause termination in casuarinas moth (*Lymantria xylina*) eggs. *Formos. Entomol* **24**, 43–52 (2004).
7. Chang, Y. & Weng, Y. Morphology, life habit, outbreak and control of casuarina tussock moth (*Lymantria xylina* Swinhoe). *Quart J Chin Forest* **18**, 29–36 (1985).
8. Zhang, J. *et al*. Reproductive and flight characteristics of *Lymantria xylina* (Lepidoptera: Erebidae) in Fuzhou, China. *Insects* **15**, 894 (2024).
9. Mastro, V. C., Munson, A. S., Wang, B., Freyman, T. & Humble, L. M. History of the Asian *Lymantria* species program: A unique pathway risk mitigation strategy. *J Integr Pest Manag* **12**, 31 (2021).
10. Zhang, J. *et al*. Evaluation of the potential flight ability of the casuarina moth, *Lymantria xylina* (Lepidoptera: Erebidae). *Insects* **15**, 506 (2024).
11. Nai, Y.-S. *et al*. Genomic sequencing and analyses of *Lymantria xylina* multiple nucleopolyhedrovirus. *BMC genomics* **11**, 116 (2010).
12. Tsay, J.-G., Lee, M.-J. & Chen, R.-S. Evaluation of *Beauveria bassiana* for controlling casuarina tussock moth (*Lymantria xylina* Swinhoe) in casuarina plantations. *Taiwan Journal for Science* **16**, 201–207 (2001).
13. Cheng, T. *et al*. Genomic adaptation to polyphagy and insecticides in a major East Asian noctuid pest. *Nat Ecol Evol* **1**, 1747–1756 (2017).
14. Xie, Q. *et al*. Identification of differentially expressed genes and proteins related to diapause in *Lymantria dispar*: Insights for the mechanism of diapause from transcriptome and proteome analyses. *PloS one* **20**, e0316065 (2025).
15. Blackburn, G. S. *et al*. Genetics of flight in spongy moths (*Lymantria dispar* ssp.): functionally integrated profiling of a complex invasive trait. *BMC genomics* **25**, 541 (2024).
16. Akhanaev, Y. *et al*. Virulence and genome analysis of baculovirus isolates from different *Lymantria dispar* populations. *Sci Rep* **15**, 28449 (2025).
17. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884-–i890 (2018).
18. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
19. 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 (2020).
20. Koren, S. *et al*. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**, 722–736 (2017).
21. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* **27**, 737–746 (2017).
22. Walker, B. J. *et al*. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one* **9**, e112963 (2014).
23. Servant, N. *et al*. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* **16**, 259 (2015).
24. Burton, J. N. *et al*. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat Biotechnol* **31**, 1119–1125 (2013).
25. Lin, Y. *et al*. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic Res* **10**, uhad127 (2023).
26. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* **35**, W265–W268 (2007).
27. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358 (2005).
28. Hoede, C. *et al*. PASTEC: an automatic transposable element classification tool. *PloS one* **9**, e91929 (2014).
29. Jurka, J. *et al*. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**, 462–467 (2005).

30. Chen, N. Using Repeat Masker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **5**, 4.10.1–4.10.14 (2004).
31. Keilwagen, J. *et al*. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res* **44**, e89–e89 (2016).
32. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**, 215–225 (2003).
33. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
34. Alioto, T., Blanco, E., Parra, G. & Guigó, R. Using geneid to identify genes. *Curr Protoc Bioinformatics* **64**, e56 (2018).
35. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**, 78–94 (1997).
36. Korf, I. Gene finding in novel genomes. *BMC bioinformatics* **5**, 21 (2004).
37. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357–360 (2015).
38. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290–295 (2015).
39. Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res* **43**, e78–e78 (2015).
40. Campbell, M. A., Haas, B. J., Hamilton, J. P., Mount, S. M. & Buell, C. R. Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. *BMC genomics* **7**, 327 (2006).
41. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7 (2008).
42. McGinnis, S. & Madden, T. L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res* **32**, W20–W25 (2004).
43. Koonin, E. V. *et al*. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol* **5**, R7 (2004).
44. Boeckmann, B. *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**, 365–370 (2003).
45. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* **18**, 366–368 (2021).
46. Conesa, A. *et al*. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
47. 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).
48. Griffiths-Jones, S. *et al*. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* **33**, D121–D124 (2005).
49. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907–915 (2019).
50. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
51. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
52. Chong, J. & Xia, J. MetaboAnalystR: an R package for flexible and reproducible analysis of metabolomics data. *Bioinformatics* **34**, 4313–4314 (2018).
53. Chen, T. *et al*. The genome sequence archive family: toward explosive data growth and diverse data types. *Genom Proteom Bioinf* **19**, 578–583 (2021).
54. Database resources of the national genomics data center, China national center for bioinformation in 2024. *Nucleic Acids Res* **52**, D18-D32 (2024).
55. *NGDC/CNCB. Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/search?searchTerm=CRA027397 (2025).
56. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP655858 (2025).
57. *NGDC/CNCB. Genome Warehouse* https://ngdc.cncb.ac.cn/gwh/Assembly/98230/show (2025).
58. Wang, R. *et al*. Chromosome-level genome assembly and genome annotation of *Lymantria xylina*. *GenBank* http://identifiers.org/ncbi/insdc:JBPSJU000000000 (2025).
59. Wang, R. *et al*. Chromosome-level genome assembly and genome annotation of *Lymantria xylina*. *Figshare* https://doi.org/10.6084/m9.figshare.29497898 (2025).
60. Wang, R. MetaboLights MTBLS13575 http://www.ebi.ac.uk/metabolights/MTBLS13575 (2025).
61. 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 (2015).
62. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**, 245 (2020).
63. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol* **20**, 277 (2019).
64. Xu, Z. *et al*. Chromosome-level genome assembly of the Asian spongy moths *Lymantria dispar asiatica*. *Sci Data* **10**, 898 (2023).
65. Wang, R. *et al*. Commands and parameters settings. *Figshare* https://doi.org/10.6084/m9.figshare.30909005 (2025).

Acknowledgements

This research was funded by the National Natural Science Foundation of China (grant numbers: 31600522) and the Natural Science Foundation of Fujian Province (grant number: 2017J0106).

Author contributions

W.R. designed this study. S.L., H.J., T.N. carried out genome sequencing. S.L., H.J., T.N., K.W., X.H., S.W. performed the data analyses and visualization. S.L., H.J., T.N. drafted the manuscript. W.R., F.Z. revised the manuscript. All authors contributed the final text of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to F.Z. or R.W.