scientific data

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of Z strain European corn borer *Ostrinia nubilalis* (Lepidoptera: Crambidae)

Xinhua Ding[1], Yue Zhang[2] ✉, Xiaowu Wang[1], Kaiyun Fu[1], Zunzun Jia[1], Zhihui Wang[3], Aerziguli Rouzi[1], Tursun Ahmat[1] & Wenchao Guo[1] ✉

European corn borer *Ostrinia nubilalis* (Hübner) is the most important pest of maize around world, and an ideal model for the sympatric host races, evolutionary and speciation research. In this study, we assembled a chromosome-level genome of Z strain *O. nubilalis* by the integrated Illumina short reads, PacBio Revio long reads, and Hi-C sequencing data. The chromosome-level genome was 480.04 Mb in total length with a contig N50 length of 16.51 Mb, which 98.59% genome anchored into 31 chromosomes. For the annotation, 1,046,695 repeat sequences in length of 212.07 Mb, 1,550 non-coding RNAs (including 1,208 tRNAs, 179 rRNAs, 62 miRNAs, 81 snRNAs, and 20 snoRNAs), and 17,145 protein-coding genes were identified. And 100% genes were functional annotated by SwissProt, NR, eggNOG, Go, and KEGG database. This genome provides a valuable genomics resource to elucidate the host plant adaptation, thermal adaptation, diapause induction, *Bacillus thuringiensis* toxin resistance, sexual communication, sympatric host races, and speciation process of *O. nubilalis*.

Background & Summary

European corn borer *Ostrinia nubilalis* (Hübner) is the most important pest around world which has successfully invaded into more than 40 countries across Africa, Asia, Europe, and North America[1–3]. It has a wide host range, feeding on more than 200 cultivated plants and weeds over Amaranthaceae, Asteraceae, Poaceae, Solanaceae, Fabaceae, Malvaceae, Cannabaceae, Rosaceae, and Salicaceae, while the most favorite host is maize[4]. Heavily attacked on maize by larvae could result in smaller plant size, fewer corn kernels, and more vulnerable during windy weather, seriously impacting on maize yield (Fig. 1). In America, it was estimated that *O. nubilalis* cost more than one billion dollars per year for control costs and yield losses[5].

Besides as widely studied agricultural pests, *O. nubilalis* is also an ideal model for the sympatric host races, evolutionary and speciation research[6–8], which consists of two incomplete reproductive isolation and discordant gene genealogies strains that differ in the male responding sex pheromone component. In the Z strain, males respond to a ratio of E-11-tetradecenyl acetate (E11-14Ac) and Z-11-tetradecenyl acetate (Z11-14Ac) as 3:97, whereas males respond to a ratio of E11-14Ac and Z11-14Ac as 99:1 in the E strain[9]. Massive studies were conducted on the genetic structure, genetic differentiation, and gene flow using AFLP (Amplified Fragment Length Polymorphism)[10], RFLP (Restriction Fragment Length Polymorphism)[11], microsatellite locus[12], and partial genes[13] as markers. Whereas, limited molecular markers might cause discordant phylogenies or genealogies when identifying the genomic location involved in reproductive barriers or adaptations. Hence, a chromosome-level reference genome will be helpful to elucidate patterns of evolutionary histories and speciation process in *O. nubilalis*.

[1]Key Laboratory of Integrated Pest Management on Crops in Northwestern Oasis, Ministry of Agriculture and Rural Affairs, National Plant Protection Scientific Observation and Experiment Station of Korla, Xinjiang Key Laboratory of Agricultural Biosafety, Institute of Plant Protection, Xinjiang Uygur Autonomous Region Academy of Agricultural Sciences, Urumqi, 830091, China. [2]Institute of Environment and Sustainable Development in Agriculture, Chinese Academy of Agricultural Sciences, Beijing, 100081, China. [3]Institution of Microbial Application, Xinjiang Academy of Agricultural Sciences, Urumqi, 830091, China. ✉e-mail: zhangyue01@caas.cn; gwc1966@163.com

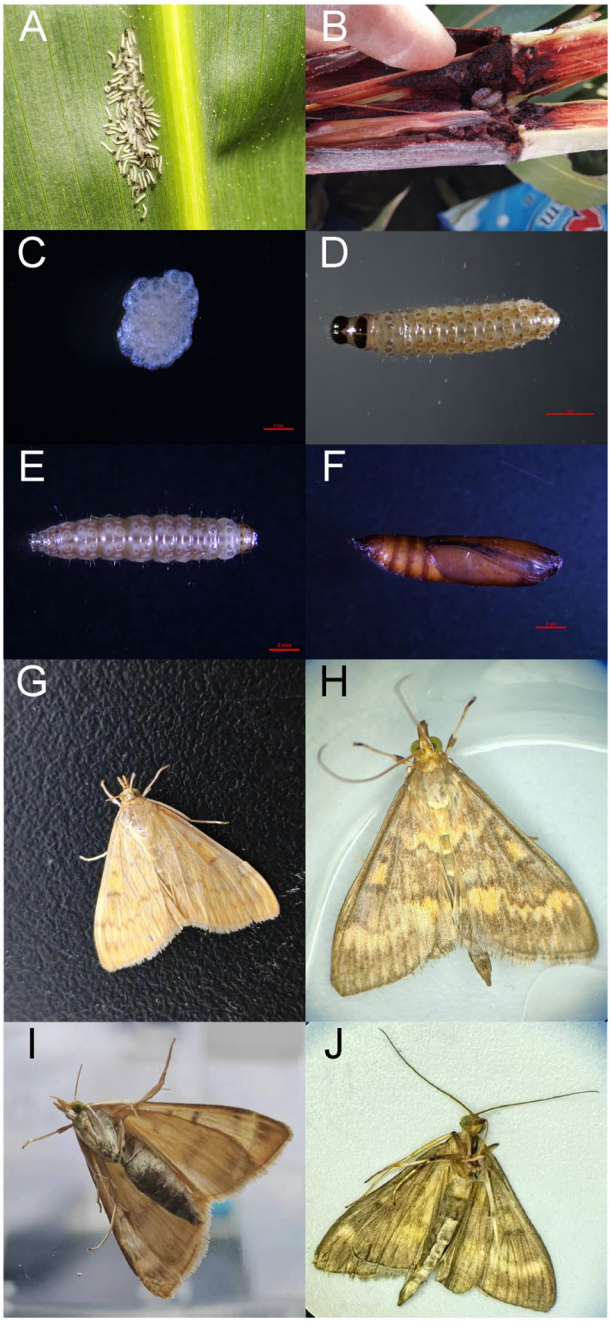

Fig. 1 Damage symptoms in maize field and morphological features of *Ostrinia nubilalis*. (**A**) larvae on a maize leaf, (**B**) larva in a maize stem, (**C**) eggs, (**D**) larva, (**E**) prepupa, (**F**) pupa, (**G**) female adult (dorsal view), (**H**) male adult (dorsal view), (**I**) female adult (ventral view), and (**J**) male adult (ventral view).

In the present study, we constructed a high-quality chromosome-level genome of Z strain *O. nubilalis* by the integrated sequencing data, including Illumina short reads, PacBio Revio long reads, and Hi-C (High-throughput Chromosomal Conformation Capture) technologies. The assembled genome was 480.04 Mb in total length with a contig N50 length of 16.51 Mb, anchoring into 31 chromosomes. Gene structure annotation identified 17,145 protein-coding genes, 1550 non-coding RNAs. This genome provides a valuable genomics resource for the research on host plant adaptation, thermal adaptation, diapause induction, *Bacillus thuringiensis* toxin resistance, sexual communication, sympatric host races, and speciation process of *O. nubilalis*.

Methods

Sample preparation. *O. nubilalis* eggs were collected from its host maize in Xinyuan County, Ili Kazakh Autonomous Prefecture, Xinjiang Uygur Autonomous Region, China (43°25′ N, 83°03′ E) in April 2023. Then the laboratory strain was established in the artificial climate chamber, fed by artificial diet[14], at condition of 28 ± 0.5 °C, 75 ± 5% relative humidity, and 16:8 light:dark photoperoid.

Type	Development stage	Usage	Sequencing platform	Number of bases (Gb)	Coverage	SRA accession number
Genome	Larvae	*De novo* assembly	PacBio Revio	15.21	31.49	SRR29025672
	Larvae	Genome survey	Illumina NovaSeq	57.54	119.13	SRR29025679
Hi-C	Larvae	Chromosome scaffolding	Illumina NovaSeq	70.26	145.51	SRR29025678
Transcriptome	All stages mixture	Annotation	PacBio Iso-seq	28.88		SRR29025671
	Eggs	Annotation	Illumina NovaSeq	8.57		SRR29025677
	Larvae	Annotation	Illumina NovaSeq	8.84		SRR29025674
	Pupae	Annotation	Illumina NovaSeq	8.91		SRR29025673
	Male adult	Annotation	Illumina NovaSeq	8.85		SRR29025675
	Female adult	Annotation	Illumina NovaSeq	9.49		SRR29025676

Table 1. Statistical characteristics of the *Ostrinia nubilalis* sequencing reads in the present study.

Genomic DNA sequencing. Genomic DNA was extracted from 5th-instar larvae using CTAB (Cetyltrimethylammonium Bromide) method[15]. The integrity of extracted genomic DNA was confirmed by Pippin Pulse using 0.7% agarose gel electrophoresis and analyzed with an Agilent 2100 Bioanalyzer (Agilent Technologies, USA), showing that the length of main band was larger than 23,130 bp. DNA concentration was measured using a Nanodrop 2000 spectrophotometer (Termo Fisher Scientific, USA) and Qubit 2.0 (Termo Fisher Scientific, USA), showing a total quantity of 11.66 μg and concentration of 259.00 ng/μL.

For Illumina sequencing, the genomic DNA was randomly fragmented using Covaris ultrasonicator (Covaris, USA) and fragments with insertion size in 350 bp were selected. Library was constructed using Illumina TruSeq Nano DNA Library Prep Kit (Illumina, USA) and sequenced on the Illumina NovaSeq 6000 platform (Illumina, USA). In total, 384,177,920 short reads in 57.63 Gb were obtained. After filtered by fastp[16], 384,177,896 reads in 57.54 Gb were finally used for genome survey and genome assembly error correction (Table 1).

For PacBio Revio sequencing, the genomic DNA was randomly fragmented using Megaruptor 3 (Diagenode, USA) and fragments with insertion size in 15 Kb were selected. Library was constructed using Pacific Biosciences SMRT bell express template prep kit 2.0 (Pacifc Biosciences, USA) and sequenced on the PacBio Revio System (Pacifc Biosciences, USA). In total, 891,291 HiFi reads in 15.21 Gb were obtained for contig-level genome assembly, with an average length of 17.06 kb and an N50 length of 17.27 kb, corresponding to 32-fold coverage of the *O. nubilalis* genome (Table 1).

Hi-C library preparation and sequencing. The larva were cut into small pieces and fixed by 2% formaldehyde. Cross-linked chromatin was digested with restriction enzyme DpnII. Through biotinylated labeled, blunt-end ligation and purified, fragmented DNA with insertion size in 350 bp was selected for library preparation. Library was sequenced on the Illumina NovaSeq 6000 platform. In total, 70.26 Gb was yielded, after filtered by fastp[16], for chromosome scaffolding (Table 1).

Transcriptome sequencing. Total RNA was isolated from eggs, larvae, pupae, female adult, and male adult using the TRIzol reagent[17], respectively. The integrity of extracted total RNA was confirmed by Pippin Pulse using 0.7% agarose gel electrophoresis and analyzed with an Agilent 2100 Bioanalyzer (Agilent Technologies, USA), showing that all the samples RIN (RNA Intergrity Number) values were higher than 8.1. RNA concentration was measured using a Nanodrop 2000 spectrophotometer (Termo Fisher Scientific, USA) and Qubit 2.0 (Termo Fisher Scientific, USA), showing all the samples quantity were higher than 10.14 μg and concentration were higher than 411.60 ng/μL.

For Illumina sequencing, five libraries (RNA extracted from five mentioned stages) were constructed using TruSeq RNA Library Preparation Kit (Illumina, USA) and sequenced on the Illumina NovaSeq 6000 platform (Illumina, USA), respectively. In total, 8.57 Gb, 8.84 Gb, 8.91 Gb, 8.85 Gb, and 9.49 Gb sequencing data yielded from five libraries after filtered by fastp[16] (Table 1).

For PacBio Iso-Seq sequencing, RNA extracted from five mentioned stages were equally mixed and then synthesized to cDNA using the Clontech SMARTer PCR cDNA Synthesis Kit (Takara Biotechnology, China). SMRTbell library was constructed with the Pacific Biosciences SMRTbell template prep kit and sequenced on the Pacific Bioscience Sequel II platform. A total of 8,879,630 subreads in 28.88 Gb were obtained for the gene structure annotation (Table 1).

Genome size and heterozygosity estimation. Illumina paired-end reads were used to preliminarily estimate the size, heterozygosity and repeat ratio of the *O. nubilalis* genome. The histogram of 19-mer frequencies were counted by Jellyfish v2.2.10[18] with parameters "count -m 19 -C -s 1 G -G 2" and visualized using GenomeScope v2.0[19] with default parameters. A total number of 15,193,094,420 19-mers were generated. With a dominant peak depth of 29, the genome size of *O. nubilalis* was estimated to be approximately 467.47 Mb, with a heterozygosity ratio of 2.34% and repeat sequence ratio of 41.85% (Fig. 2A).

Genome assembly. For contig-level genome assembly, a draft genome was assembled by Hifiasm v0.19.5[20] with default parameters using PacBio Revio sequences. Then, the draft genome was further filtered by the following step: (1) removing haplotigs and contigs overlaps based on read dept by Purge_dups v1.2.5[21] with

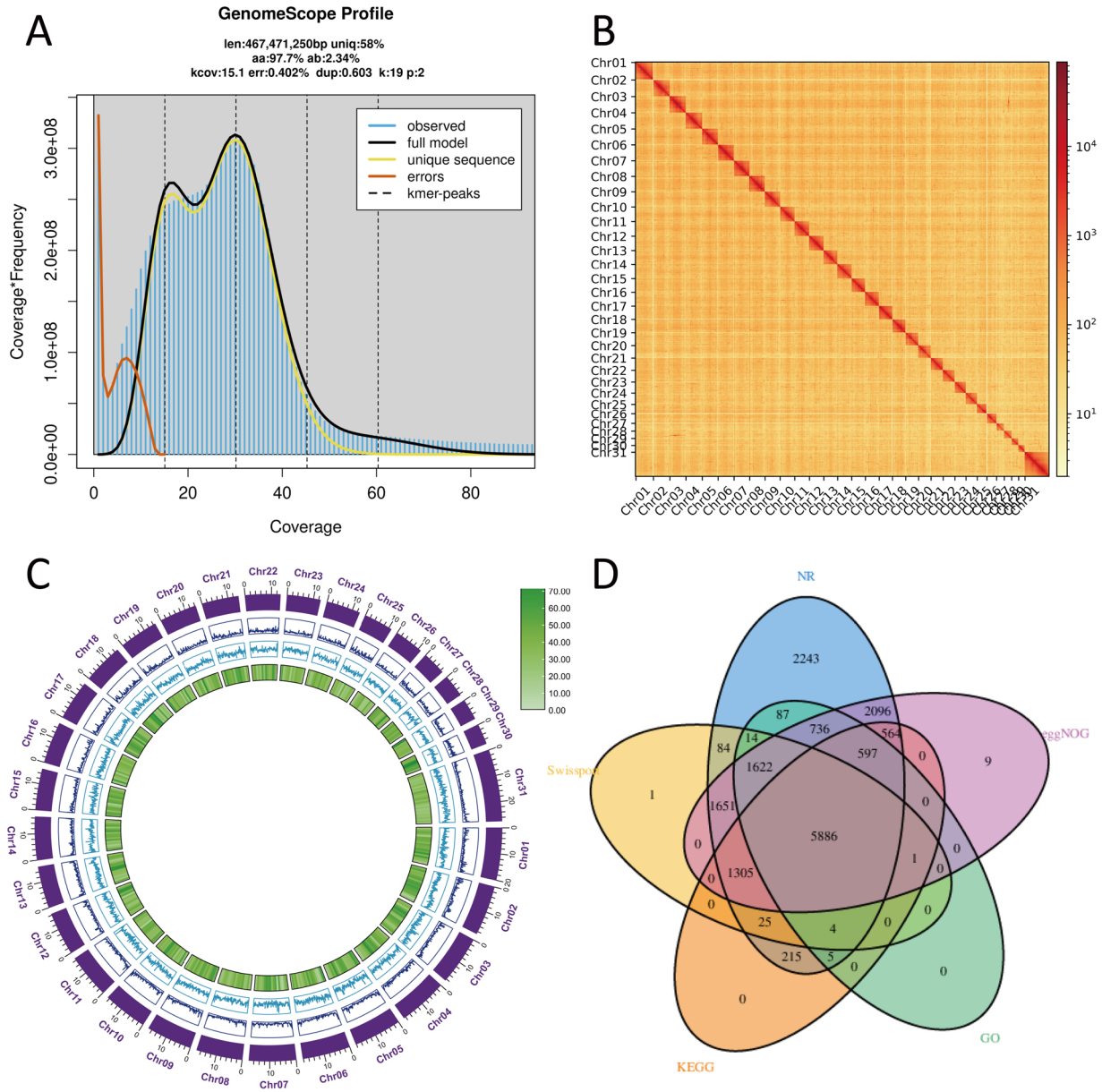

Fig. 2 Genomic characteristics of *Ostrinia nubilalis*. (**A**) Genome size estimation by 19 K-mer frequency distribution analysis based on Illumina short paired-end sequencing data. (**B**) Heatmap of genome-wide all-by-all Hi-C interaction within 31 chromosomes. The colour bar illuminates the Hi-C interactive intensity ranging from low (white) to high (dark red) in the plot. (**C**) Genomic landscape of the 31 chromosomes. Tracks from outer to inner represent length of chromosomes at the Mb scale, CG content per 100 kb, repeat sequences ratio per 100 kb, and protein-coding genes density per 100 kb, respectively. (**D**) Venn diagram of shared and unique functional annotated protein-coding genes by five databases: SwissProt (Swiss Institute of Bioinformatics and Protein Information Resource), NR (Non-Reduntant Protein Database), eggNOG (Evolutionary Genealogy Of Genes: Non-Supervised Orthologous Groups), Go (Gene Ontology), and KEGG (Kyoto Encyclopedia of Gene and Genomes).

parameters "-2 -T changed.cutoffs", (2) removing bacteria, viruses sequences by blast v2.12.0[22] against NT database with parameters "-evalue 1e-5", (3) filtering contigs, which average depth lower than 5-fold, by aligning Revio sequences against draft genome using minimap2 v2.26[23]. Eventually, a 483.01 Mb in length contig-level genome, consisting by 119 contigs ranging from 20.86 Kb to 18.90 Mb, was generated with a contig N50 length of 9.43 Mb (Table 2).

For chromosome-level scaffolding, Hi-C sequencing reads were mapped to the contig-level genome using HiCPRO v2.11.4[24] and 20,776,488 (8.97% of total reads) valid interaction pairs were obtained for the further analysis. Next, the 3D-DNA pipeline[25] was used to order, orient, and cluster the contigs. Eventually, a chromosome-level genome in length of 480.04 Mb with a scaffold N50 of 16.51 Mb was generated, with 98.59%

Genome assembly statistics	Contig-level	Chromosome-level
Sequences number	119	48
Total length (bp)	483,012,570	480,042,831
GC (%)	37.7	37.7
Minimum length (bp)	20,864	10,000
N50 length (bp)	9,431,604 (n = 19)	16,509,001 (n = 13)
N90 length (bp)	2,356,287 (n = 59)	11,010,267 (n = 27)
Maximum length (bp)	18,901,614	27,381,392

Table 2. Summary of *Ostrinia nubilalis* genome assembly.

Chromosome	Contigs number	Chromosome length (bp)
Chr01	7	20,008,304
Chr02	1	18,901,614
Chr03	3	18,567,400
Chr04	6	18,659,698
Chr05	2	18,128,657
Chr06	5	18,178,055
Chr07	2	17,864,840
Chr08	5	17,573,208
Chr09	4	17,323,471
Chr10	4	16,651,201
Chr11	5	16,762,974
Chr12	1	16,409,655
Chr13	2	15,884,356
Chr14	1	16,509,001
Chr15	1	15,677,521
Chr16	4	15,788,599
Chr17	4	15,094,708
Chr18	1	15,194,996
Chr19	3	14,667,924
Chr20	3	14,253,820
Chr21	3	13,942,070
Chr22	5	13,324,401
Chr23	2	12,661,923
Chr24	3	12,872,566
Chr25	3	11,010,267
Chr26	1	11,059,469
Chr27	2	8,825,063
Chr28	5	8,439,178
Chr29	5	7,937,514
Chr30	4	7,739,194
Chr31	8	2,738,1392

Table 3. Summary of chromosomes in the genome of *Ostrinia nubilalis*.

of the genome anchored to 31 chromosomes (Tables 2, 3). The well-distinguished interaction matrix heatmap was visualized by HiCExplorer[26] (Fig. 2B).

Repeat sequences annotation. Repeat sequences were identified based on two strategies: homology and *de novo*. Firstly, MITE-Hunter-11-2011[27] was used to identify MITEs (miniature inverted-repeat transposable elements) to build MITEs library. Secondly, LTRharvest v1.6.5[28] and LTR Finder v1.0.7[29] were used to identify LTRs (long terminal repeats) and then non-redundant accurate LTR-RT library were generate by LTR retriever v2.9.8[30]. Thirdly, RepeatMasker v1.323[31] was used to identify repeat sequences in the genome against the Repbase[32]. Fourthly, RepeatMasker v1.323[31] masked the known repeat sequences from the above three steps. Fifthly, RepeatModeler open-1.0.8[33] was used to identify TEs (transposable elements) in the genome. Eventually, combined all the identification results, 1,046,695 repeat sequences in length of 212.07 Mb (44.18% of the genome) were identified in *O. nubilalis*, including retrotransposon (114.02 Mb in length, accounting for 53.76%), DNA

Classification	Super family	Classification	Number	Length (bp)	Percentage in the genome
Retrotransposon	LTR	LTR	201,723	30,926,826	6.44%
		LTR/Copia	36,159	9,374,834	1.95%
		LTR/DIRS	1,308	932,552	0.19%
		LTR/Gypsy	161,807	58,071,074	12.10%
		LTR/Pao	60	5,491	0.00%
	Non-LTR	LINE	831	585,359	0.12%
		LINE/CR1	4,058	961,248	0.20%
		LINE/CR1-Zenon	2,564	474,393	0.10%
		LINE/CRE	785	195,823	0.04%
		LINE/Dong-R4	324	265,184	0.06%
		LINE/I	2,347	1,526,266	0.32%
		LINE/I-Jockey	2,116	1,040,274	0.22%
		LINE/L1-Tx1	79	11,195	0.00%
		LINE/L2	12,947	3,077,188	0.64%
		LINE/Penelope	2,503	164,718	0.03%
		LINE/Proto2	667	183,425	0.04%
		LINE/R1	7,510	2,742,527	0.57%
		LINE/R1-LOA	1,559	432,388	0.09%
		LINE/R2	1	87	0.00%
		LINE/RTE	138	21,904	0.01%
		LINE/RTE-BovB	4,443	716,524	0.15%
		LINE/RTE-RTE	27,926	1,979,237	0.41%
		SINE	167	17,663	0.00%
		SINE/5S	5,075	278,471	0.06%
		SINE/B2	503	29,871	0.01%
		SINE/U	39	3,656	0.00%
		SINE/tRNA-Deu-L2	1	34	0.00%
DNA Transposon	DNA-TE	DNA	424,297	78,469,562	16.35%
		DNA/Academ-1	645	335,333	0.07%
		DNA/CMC-Chapaev-3	78	24,854	0.01%
		DNA/CMC-Transib	629	123,730	0.03%
		DNA/Ginger-2	136	16,597	0.00%
		DNA/MULE-MuDR	101	8,379	0.00%
		DNA/Maverick	1,005	669,818	0.14%
		DNA/P	1,755	651,336	0.14%
		DNA/PIF-Harbinger	1,091	664,801	0.14%
		DNA/PiggyBac	243	40,547	0.01%
		DNA/Sola-1	607	205,945	0.04%
		DNA/TcMar-Fot1	181	109,471	0.02%
		DNA/TcMar-Mariner	454	191,016	0.04%
		DNA/TcMar-Pogo	1	339	0.00%
		DNA/TcMar-Tc1	708	413,212	0.09%
		DNA/TcMar-Tigger	57	6,613	0.00%
		DNA/hAT	475	55,483	0.01%
		DNA/hAT-Ac	411	111,658	0.02%
		DNA/hAT-Blackjack	224	27,458	0.01%
		DNA/hAT-Charlie	1,091	132,415	0.03%
		DNA/hAT-Tag1	145	32,865	0.01%
		DNA/hAT-Tip100	2,600	354,528	0.07%
		DNA/hAT-hAT5	154	25,670	0.01%
		DNA/hAT-hATm	73	41,006	0.01%
		DNA/hAT-hATx	322	67,092	0.01%
		DNA/hAT-hobo	129	16,916	0.00%
	Helitron	RC/Helitron	10,495	1,159,337	0.24%
Others	Satellite	Satellite	421	592,081	0.12%
	Unknown	Unknown	120,526	13,501,949	2.81%
Total			1,046,695	212,068,261	44.18%

Table 4. Summary of repeats elements identified in the genome of *Ostrinia nubilalis*. Note: LTR, Long Terminal Repeat; LINE, Long Interspersed Nuclear Elements; SINEs, Short Interspersed Nuclear Elements; TE, Transposable Element; RC, Rolling-circle Eukaryotic Transposons.

Classification	Number	Average length (bp)	Total length (bp)
rRNA	179	1,405	251,576
miRNA	62	78	4,876
tRNA	1,208	72	87,712
snRNA	81	144	11,669
snoRNA	20	143	2,868

Table 5. Summary of non-coding RNAs identified in the genome of *Ostrinia nubilalis*.

Strategies		Number of genes	Mean CDS length (bp)	Exons per transcript	Mean exon length (bp)	Mean intron length (bp)
De novo	GlimmerHMM	37,624	833	4.9	168	2,729
	AUGUSTUS	22,697	1,464	6.3	231	1,916
	SNAP	12,402	1,871	10.9	171	6,749
	GeneMark-ET	21,100	1,272	6.3	201	984
Transcripts		24,064	1,347	6.8	690	2,548
Homology	*Cnaphalocrocis medinalis*	10,888	1,352	5.4	249	1,275
	Chilo suppressalis	9,518	1,406	6.3	222	1,303
	Ostrinia furnacalis	13,767	1,442	5.5	264	1,299
	Heortia vitessoides	14,798	1,263	5.7	223	1,214
	Ostrinia nubilalis	15,432	1,119	4.5	250	1,272
	Conogethes punctiferalis	20,509	1,410	4.6	308	1,195

Table 6. Statistics of gene structure identified by three strategies in the genome of *Ostrinia nubilalis*.

Number of genes	17,145
Total genic length	275,019,141 bp
Mean gene length	16,040 bp
Number of transcripts	19,369
Transcripts per gene	1.1
Total transcript length	53,773,662 bp
Mean transcript length	2,776 bp
Number of exons	144,381
Exons per transcript	7.5
Mean exon length	372 bp
Number of coding exons	140,101
Number of introns	125,012
Mean intron length	2,341 bp
Total cds length	34,167,466 bp
Mean CDS length	1,764 bp

Table 7. Summary of gene structure identification combined by EVM in the genome of *Ostrinia nubilalis*.

transposon (83.96 Mb in length, accounting for 39.59%), satellite (0.59 Mb in length, accounting for 0.28%), and unknown (13.50 Mb in length, accounting for 6.37%) (Table 4).

Non-coding RNA annotation. tRNAscan-SE v1.3.143[34] was used to identified tRNAs (Transfer RNAs). And Rfam[35] was used to identified rRNAs (Ribosomal RNAs), miRNAs (Micro RNAs), snRNAs (Small Nuclear RNAs), and snoRNAs (Small Nucleolar RNAs). In total, 1,550 non-coding RNAs were identified in *O. nubilalis*, including 1,208 tRNAs, 179 rRNAs, 62 miRNAs, 81 snRNAs, and 20 snoRNAs (Table 5).

Protein-coding genes prediction and functional annotation. Three strategies were used to predict gene structure from the repeat sequence masked genome. Firstly, for the transcriptome-based strategy, HISAT2 v2.2.1[36] was used to mapped Illumina RNA-seq data to the genome with parameter "-q–dta-cufflinks -x" and then Cufflinks v2.2.1[37] was used to assemble into transcripts with parameter "-p 8 -u–upper quartile-norm". Meanwhile, SMRTLink v13.1 software (available at https://www.pacb.com/smrt-link/) was used to generate full-length transcripts from PacBio Iso-Seq data. Then PASA v2.1 pipeline[38] was used to predict ORFs (open reading frames) with parameters "-c alignAssembly.config -C -R–ALIGNERS blat,gmap–CPU 32–stringent_alignment_overlap 30.0". Secondly, for the *de novo*-based strategy, Augustus v3.5.0[39] (with parameter

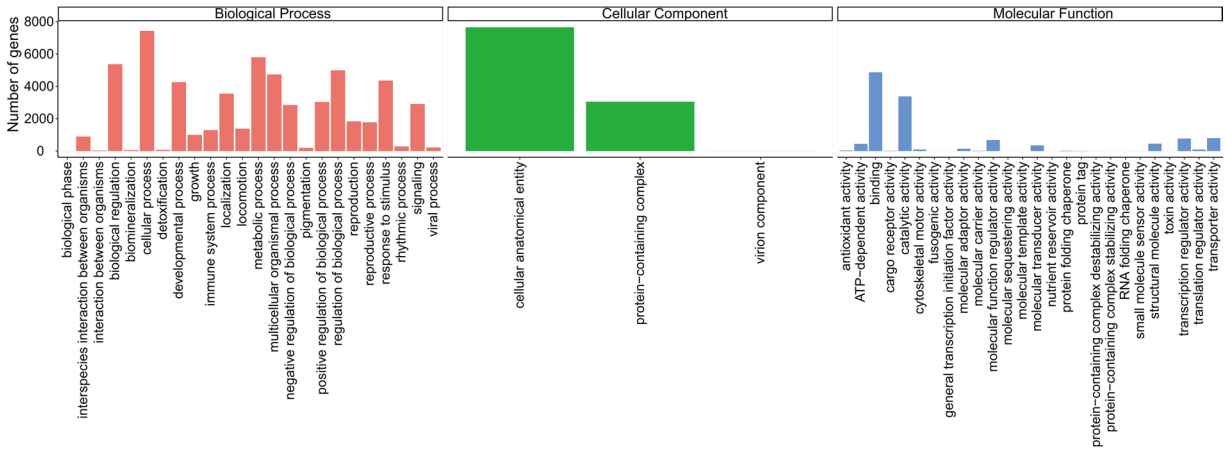

Fig. 3 GO annotation statistics of protein-coding genes in *Ostrinia nubilalis* genome. The horizontal axis shows the GO classification types, and the vertical axis represents the number of annotated protein-coding genes.

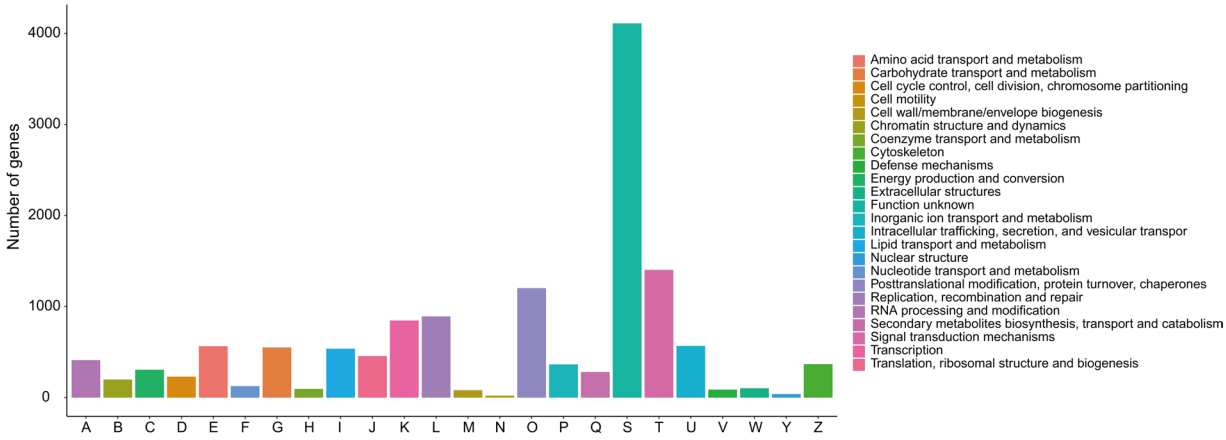

Fig. 4 COG annotation statistics of protein-coding genes in *Ostrinia nubilalis* genome. The horizontal axis shows the COG classification types, and the vertical axis represents the number of annotated protein-coding genes.

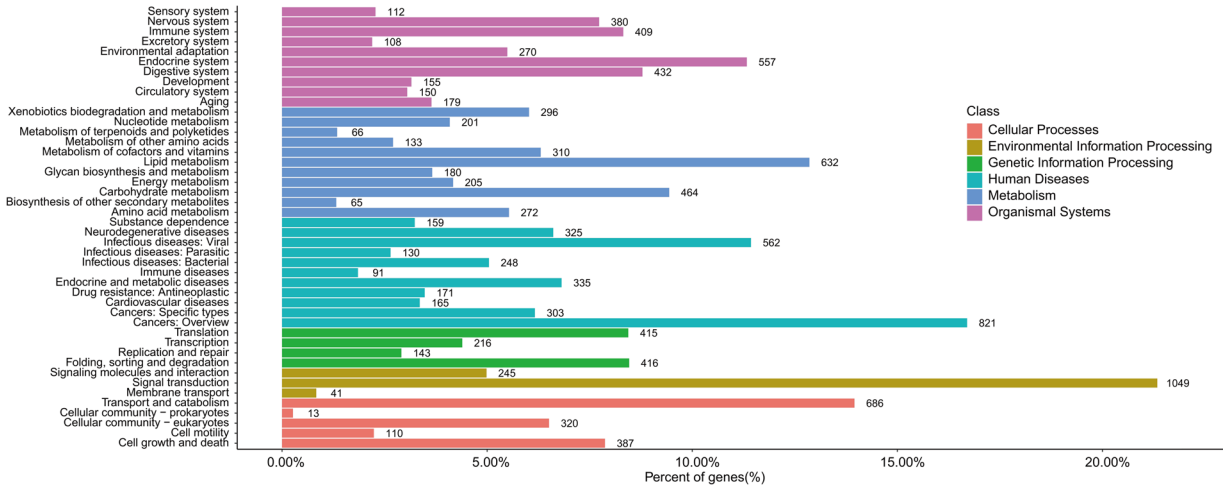

Fig. 5 KEGG pathway annotation statistics of protein-coding genes in *Ostrinia nubilalis* genome. The horizontal axis shows the number of annotated protein-coding genes, and the vertical axis represents the KEGG pathway classification types.

	Number	Percentage
Protein-coding genes	17,145	100%
Annotated genes	17,145	100%
Annotated by NR	17,134	99.94%
Annotated by SwissProt	10,593	61.78%
Annotated by KEGG	8,602	50.17%
Annotated by GO	8,952	52.21%
Annotated by eggNOG	14,467	84.38%
Not Annotated	0	0%

Table 8. Summary of protein-coding genes functional annotation in the genome of *Ostrinia nubilalis*.

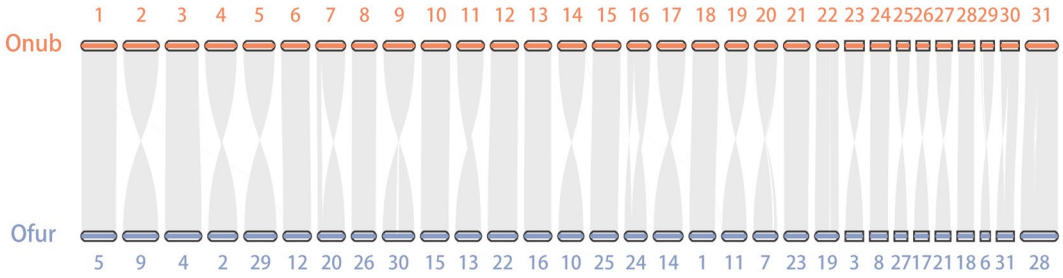

Fig. 6 Whole-genome collinearity between *Ostrinia nubilalis* (Onub) and *Ostrinia furnacalis* (Ofur). Links represent the syntenic blocks identified by jcvi.

"–species–alternatives-from-evidence–alternatives-from-sampling–allow_hinted_splicesites = atac"), SNAP v38926[40] (with default parameters), GlimmerHMM v.3.0425[41] (with default parameters), and GeneMark-ET v4.32[42] (with parameter "–ET–cores 48–max_gap 3000") were used to identify protein-coding genes. Thirdly, for the homology-based strategy, GeMoMa v1.6[43] with default parameters were used to predict gene structure by comparing with six genome-annotated close species *Cnaphalocrocis medinalis*[44], *Chilo suppressalis*[45], *Ostrinia furnacalis*[46], *Heortia vitessoides*[47], *O. nubilalis*[48], and *Conogethes punctiferalis*[49]. Finally, EVM v1.11[50] integrated all the results into a non-redundant gene set and PASA v2.1[38] annotated the UTRs (untranslated regions) and alternative splicing variants. In total, 17,145 protein-coding genes were identified with an mean length of 16,040 bp, mean transcript length of 2,776 bp, mean exon length of 372 bp, mean intron length of 2,341 bp, and mean CDS length of 1,764 bp (Tables 6, 7).

Protein-coding genes were functional annotated using Blast[22], with an E-value cut of 1e-5, against SwissProt[51], NR, eggNOG[52], KEGG[53], and GO[54] database (Figs. 3–5). Hmmer v3.1b160[55] was used to predict the protein domains against the Pfam Database[56]. In sum, all 17,145 protein-coding genes were functional annotated by at least one database (Fig. 2D, Table 8).

Chromosome-level genomic landscape visualization. Bedtools v2.30.0[57] was used to split each chromosome into non-overlapping 100 kb sliding windows. Then seqtk-1.4[58] calculated the GC content of every window. Bedtools v2.30.0[57] calculated the repeat sequences content and genes density of every window. The chromosome-level genomic landscape circos map was visualization in TBtools-II v2.088[59] (Fig. 2C).

Chromosome-level synteny. The syntenic blocks between *O. nubilalis* and *O. furnacalis* were identified by MCScanX[60] with default parameters and visualized by jcvi v1.4.12[61] with parameter "–minspan = 30". The high degree shared syntenic blocks revealed a very close relationship between them (Fig. 6).

Data Records

The *Ostrinia nubilalis* genome project was deposited into the NCBI (National Center for Biotechnology Information) under BioProject accession number PRJNA1111073.

The genomic Illumina sequences are available at SRA (Sequence Read Archive) under accession number SRR29025679[62].

The PacBio Revio sequences are available at SRA under accession number SRR29025672[63].

The Hi-C sequences are available at SRA under accession number SRR29025678[64].

The transcriptomic Illumina sequences are available at SRA under accession numbers SRR29025673-SRR29025677[65–69].

The PacBio Iso-Seq sequences are available at SRA under accession number SRR29025671[70].

The chromosome-level assembly genome sequences are available at GenBank under the WGS accession JBDNCF000000000[71].

The genome annotation results are available in Figshare database https://doi.org/10.6084/m9.figshare.25809568.v1[72].

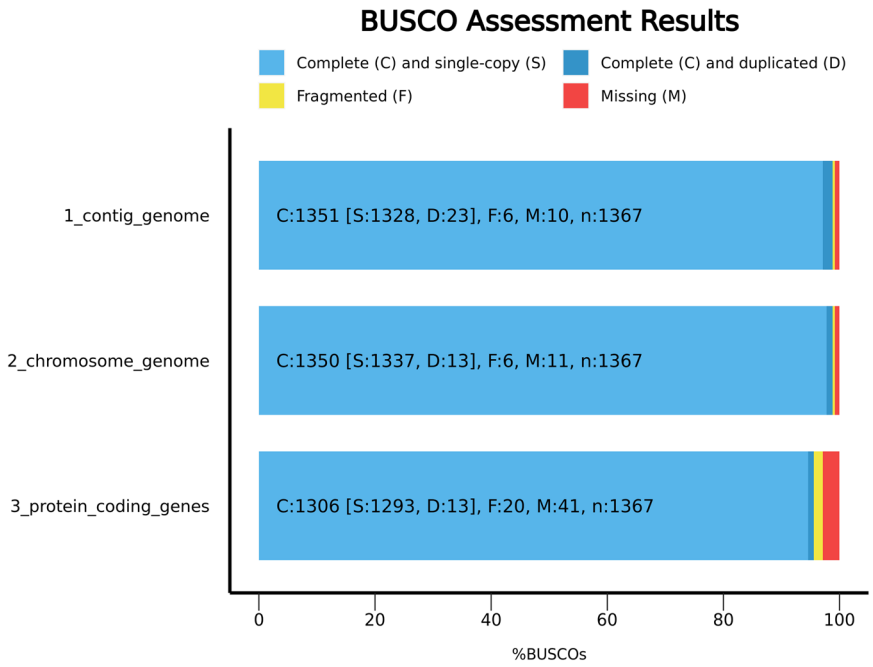

Fig. 7 Completeness evaluation of contig-level genome assembly, chromosome-level genome assembly, and predicted protein-coding genes of *Ostrinia nubilalis* by BUSCO based on insecta-odb10 database. C: the number of complete genes, S: the number of complete and single-copy genes, D: the number of complete and duplicated genes, F: the number of incomplete genes, and M: the number of missing genes.

Technical Validation

Pheromone strain determination. Sex pheromone-producing gland was isolated from 2-day-old virgin female after emergency with distilled n-hexane. The components of the sex pheromone were identified using The Shimadzu QP-2020 NX GC-MS (Gas Chromatography-Mass Spectrometry) System (Kyoto, Japan) with DB-5 capillary column (30 m × 0.25 μm × 0.25 mm) (Agilent Technologies, USA) under the following conditions: carrier gas nitrogen at a flow rate of 3 mL/min, ionization voltage 70 eV, temperature of theion source 230 °C, interface temperature 250 °C, scanning mass range of m/z 40–550. The pheromone strain was determined by the ratio of Z11-14Ac/E11-14Ac. The result clearly showed that the European corn borer population used in the present study was Z strain[72].

Evaluating the accuracy of the genome assembly. PacBio Revio sequences were mapped to the draft genome by minimap2 v2.26[23], showing that 99.82% reads were mapped to the draft genome. Mosdepth v0.3.1[73] calculated that the average depth reached 26.81-fold with 100% coverage rate of the genome, indicating an accuracy contig-level genome assembly.

Evaluating the completeness of the genome assembly and annotation. BUSCO (Benchmarking Universal Single-Copy Orthologs) v5.1.2[74] was used to evaluate the completeness of the contig-level genome, chromosome-level genome, and predicted protein-coding genes based on the insecta_odb10 database (1,367 genes). The results showed that 98.8%, 98.8%, and 95.6% complete genes found in the contig-level genome, chromosome-level genome, and predicted protein-coding genes, respectively (Fig. 7).

In sum, the evaluation results demonstrated a high-quality chromosome-level genome of *O. nubilalis* with annotated gene structure was obtained in the present study.

Code availability

No specific scripts or code were used in the present study. The version and parameters of software/pipelines used were detailed described in Methods section.

Received: 21 May 2024; Accepted: 11 February 2025;
Published online: 01 March 2025

References

1. Mutuura, A. & Munroe, E. Taxonomy and distribution of the European corn borer and allied species: genus *Ostrinia* (Lepidoptera: Pyralidae). *The Memoirs of the Entomological Society of Canada* **102**(S71), 1–112 (1970).
2. Schmutzenhofer, H., Mielke, M. E., Luo, Y., Ostry, M. E. & Wen. J. Field Guide/Manual on the Identification and Management of Poplar Pests and Diseases in the Area of the "Three North 009 Project" (North-Eastern China) (China Forestry Publishing House, 1996).
3. EPPO. EPPO Global database. In: *EPPO Global database*. https://gd.eppo.int/ (2023).

4. Ponsard, S. *et al*. Carbon stable isotopes: a tool for studying the mating, oviposition, and spatial distribution of races of European corn borer, *Ostrinia nubilalis*, among host plants in the field. *Can. J. Zool.* **82**(7), 1177–1185 (2004).
5. Mason, C. E. *et al*. European corn borer ecology and management (North Central Regional Extension Publication, 1996).
6. Malausa, T. *et al*. Assortative mating in sympatric host races of the European corn borer. *Science* **308**(5719), 258–260 (2005).
7. Dopman, E. B., Pérez, L., Bogdanowicz, S. M. & Harrison, R. G. Consequences of reproductive barriers for genealogical discordance in the European corn borer. *P. Natl. Acad. Sci. USA* **102**(41), 14706–14711 (2005).
8. Kozak, G. M. *et al*. Genomic basis of circannual rhythm in the European corn borer moth. *Curr. Biol.* **29**, 3501–3509 (2019).
9. Klun, J. A. *et al*. Insect sex pheromones: minor amount of opposite geometrical isomer critical to attraction. *Science* **181**(4100), 661–663 (1973).
10. Coates, B. S. *et al*. Frequency of hybridization between *Ostrinia nubilalis* E-and Z-pheromone races in regions of sympatry within the United States. *Ecol. Evol.* **3**, 2459–2470 (2013).
11. Koutroumpa, F. A., Groot, A. T., Dekker, T. & Heckel, D. G. Genetic mapping of male pheromone response in the European corn borer identifies candidate genes regulating neurogenesis. *P Natl. Acad. Sci. USA* **113**, E6401–E6408 (2016).
12. Li, J. *et al*. The genetic structure of Asian corn borer, *Ostrinia furnacalis*, populations in China: haplotype variance in northern populations and potential impact on management of resistance to transgenic maize. *J. Hered.* **105**(5), 642–655 (2014).
13. Piwczyński, M. *et al*. High regional genetic diversity and lack of host-specificity in *Ostrinia nubilalis* (Lepidoptera: Crambidae) as revealed by mtDNA variation. *B. Entomol. Res.* **106**, 512–521 (2016).
14. Qiao, L., Zheng, J., Cheng, W. & Li, Y. Impact of 4 different artificial fodders on life span of Asian corn borer *Ostrinia furnacalis* (Guenee). *Journal of Northwest A & F University (Nat.Sci.Ed.)* **36**(5), 109–112 (2008).
15. Allen, G. C., Flores-Vergara, M. A., Krasynanski, S., Kumar, S. & Tompson, W. F. A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. *Nat. Protoc.* **1**, 2320–2325 (2006).
16. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
17. Rio, D. C., Ares, M., Hannon, G. J. & Nilsen, T. W. Purification of RNA using TRIzol (TRI reagent). *Cold Spring Harbor Protocols* **6**, pdb–prot5439 (2010).
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. Vurture, G. W. *et al*. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
20. 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**(2), 170–175 (2021).
21. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
22. McGinnis, S. & Madden, T. L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* **32**(suppl_2), W20–W25 (2004).
23. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
24. Servantm, N. *et al*. HiC-Pro: An optimized and flexible pipeline for Hi-C processing. *Genome Biol.* **16**, 259 (2015).
25. Dudchenko, O. *et al*. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome length scaffolds. *Science* **356**(6333), 92–95 (2017).
26. Wolff, J. *et al*. Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. *Nucleic Acids Res.* **48**, W177–W184 (2020).
27. Han, Y. & Wessler, S. R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* **38**(22), e199 (2010).
28. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 1–14 (2008).
29. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–268 (2007).
30. Ou, S. & Jiang, N. LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**(2), 1410–1422 (2018).
31. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform.* **25**, 4–10 (2009).
32. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**, 11 (2015).
33. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2020).
34. 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).
35. Griffths-Jones, S. *et al*. Rfam: an RNA family database. *Nucleic Acids Res.* **31**(1), 439–441 (2003).
36. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**(4), 357–360 (2015).
37. Trapnell, C. *et al*. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**(5), 511–515 (2010).
38. Haas, B. J. *et al*. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
39. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**(suppl_2), W309–W312 (2004).
40. Leskovec, J. & Sosič, R. Snap: A general-purpose network analysis and graph-mining library. *ACM T. Intel. Syst. Tec.* **8**(1), 1–20 (2016).
41. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. *Bioinformatics* **20**(16), 2878–2879 (2004).
42. Lomsadze, A., Burns, P. D. & Borodovsky, M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res.* **42**(15), e119–e119 (2014).
43. Keilwagen, J., Hartung, F. & Grau, J. GeMoMa: Homology-based gene prediction utilizing intron position conservation and RNAseq data. In *Gene prediction*, Kollmar, M. ed, pp. 161-177 (New York, USA: Springer, 2019).
44. Zhao, X. *et al*. A chromosome-level genome assembly of rice leaffolder, *Cnaphalocrocis medinalis*. *Mol. Ecol. Resour.* **21**(2), 561–572 (2021).
45. 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**(1), 268–282 (2020).
46. Peng, Y. *et al*. Population genomics provide insights into the evolution and adaptation of the Asia corn borer. *Mol. Biol. Evol.* **40**(5), msad112 (2023).
47. Law, S. T. *et al*. Chromosomal-level reference genome of the moth *Heortia vitessoides* (Lepidoptera: Crambidae), a major pest of agarwood-producing trees. *Genomics* **114**(4), 110440 (2022).
48. Kozak, G. M. *et al*. Genomic basis of circannual rhythm in the European corn borer moth. *Curr. Biol.* **29**(20), 3501–3509 (2019).

49. Gao, B. *et al*. Chromosome genome assembly and whole genome sequencing of 110 individuals of *Conogethes punctiferalis* (Guenée). *Sci. Data* **10**(1), 805 (2023).
50. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**(1), 1–22 (2008).
51. Boeckmann, B. *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).
52. 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 (2019).
53. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.* **49**, D545–D551 (2021).
54. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
55. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**(10), e1002195 (2011).
56. Finn, R. D. *et al*. Pfam: the protein families database. *Nucleic Acids Res.* **42**(Database issue), 222–30 (2014).
57. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**(6), 841–842 (2010).
58. Li, H. Seqtk Toolkit for processing sequences in FASTA/Q formats. *GitHub* **767**, 69 (2012).
59. Chen, C. *et al*. TBtools-II: A "one for all, all for one" bioinformatics platform for biological big-data mining. *Mol. Plant.* **16**, 1733–1742 (2023).
60. Wang, Y. P. *et al*. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49–e49 (2012).
61. Tang, H. *et al*. Synteny and collinearity in plant genomes. *Science* **320**(5875), 486–488 (2008).
62. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR29025679 (2024).
63. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR29025672 (2024).
64. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR29025678 (2024).
65. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR29025673 (2024).
66. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR29025674 (2024).
67. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR29025675 (2024).
68. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR29025676 (2024).
69. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR29025677 (2024).
70. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR29025671 (2024).
71. Zhang, Y. *et al*. *GenBank* https://identifiers.org/ncbi/insdc:JBDNCF000000000 (2024).
72. Zhang, Y. *et al*. Chromosome-level genome assembly of Z strain European corn borer *Ostrinia nubilalis* (Lepidoptera: Crambidae). figshare. *Dataset*. https://doi.org/10.6084/m9.figshare.25809568.v1 (2024).
73. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**(5), 867–868 (2018).
74. Seppey, M., Manni, M. & Zdobnov, E.M. BUSCO: assessing genome assembly and annotation completeness. In *Gene prediction*, Kollmar, M. ed. pp. 227–245 (New York, USA: Springer, 2019).

Acknowledgements

The work was supported by the Xinjiang Major Science and Technology Projects (Research, development, and demonstration of key technologies for the green control of major pests on special and superiority crops in Xinjiang, 2023A02009).

Author contributions

W.G. and Y.Z. designed the study; X.D., A.R., and T.A. collected and reared samples in laboratory for sequencing; X.D., Y.Z., X.W., and K.F. assembled and annotated the genome; Z.J. and Z.W. took photographs of damage symptoms and morphological features; X.D. and Y.Z. wrote the draft manuscript. X.W., K.F., Z.J., Z.W., A.R., T.A., and W.G. revised the manuscript. All authors approved the final manuscript for submission.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.Z. or W.G.