Received: 17 June 2020 | Revised: 21 August 2020 | Accepted: 1 October 2020

DOI: 10.1111/1755-0998.13274

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

A chromosome-level genome assembly of rice leaffolder, *Cnaphalocrocis medinalis*

Xianxin Zhao[1] | Hongxing Xu[2] | Kang He[1] | Zhenmin Shi[1] | Xi Chen[1] | Xinhai Ye[1] | Yang Mei[1] | Yajun Yang[2] | Meizhen Li[1] | Libin Gao[1] | Le Xu[1] | Huamei Xiao[1,3] | Ying Liu[4] | Zhongxian Lu[2] | Fei Li[1]

[1]State Key Laboratory of Rice Biology & Ministry of Agricultural and Rural Affairs, Key Laboratory of Molecular Biology of Crop Pathogens and Insect Pests, Institute of Insect Sciences, Zhejiang University, Hangzhou, China

[2]State Key Laboratory for Managing Biotic and Chemical Treats to the Quality and Safety of Agroproducts, Institute of Plant Protection and Microbiology, Zhejiang Academy of Agricultural Sciences, Hangzhou, China

[3]College of Life Sciences and Resource Environment, Key Laboratory of Crop Growth and Development Regulation of Jiangxi Province, Yichun University, Yichun, China

[4]Agriculture Environment and Resources Institute, Yunnan Academy of Agricultural Sciences, Kunming, China

Correspondence
Fei Li, State Key Laboratory of Rice Biology & Ministry of Agricultural and Rural Affairs, Key Laboratory of Molecular Biology of Crop Pathogens and Insect Pests, Institute of Insect Sciences, Zhejiang University, Hangzhou 310058, China.
Email: lifei18@zju.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Number: 31772238; Earmarked Fund for China Agriculture Research System, Grant/Award Number: CARS-01-36; The State Key Laboratory for Managing Biotic, Chemical Treats to the Quality and Safety of Agro-products, Grant/Award Number: 2010DS700124-ZZ2007

Abstract

The rice leaffolder, *Cnaphalocrocis medinalis* Guenée (Crambidae, Lepidoptera), is an important agricultural pest that causes serious losses to rice production in rice-growing regions with high humidity and temperature. However, a lack of genomic resources limits in-depth understanding of its biological characteristics and ecological adaptation. Here, we sequenced the genome of rice leaffolder using the Illumina and PacBio platforms, yielding a genome assembly of 528.3 Mb with a contig N50 of 524.6 kb. A high percentage (96.4%) of Benchmarking Universal Single-Copy Orthologs (BUSCOs) were successfully detected, suggesting high-level completeness of the genome assembly. In total, 39.5% of the genome consists of repeat sequences and 15,045 protein-coding genes were annotated. Comparative phylogenomic analysis showed that some gene families associated with hormone biosynthesis expanded in rice leaffolder. Next, we used the Hi-C technique to produce a chromosome-level genome assembly with a scaffold N50 of 16.1 Mb by anchoring 3,248 scaffolds to 31 chromosomes. The rice leaffolder genome showed high chromosomal synteny with the genome of four other lepidopteran insects. By comparing coverage ratios from the genome resequencing of male and female pupae, we identified near intact Z and W chromosomes. The W chromosome is estimated as 20.75 Mb, which is the most complete known W chromosome in Lepidoptera. The protein-coding genes on the W chromosome were significantly enriched in metabolic pathways. In all, the high-quality genome assembly and the near-intact W chromosome of rice leaffolder should be a useful resource for the fields of insect migration, chromosome evolution and pest control.

KEYWORDS

chromosome-level genome, *Cnaphalocrocis medinalis*, comparative genomics, genome synteny, migration, rice leaffolder

Xianxin Zhao and Hongxing Xu are contributed equally to this work.

1 | INTRODUCTION

The rice leaffolder, *Cnaphalocrocis medinalis*, a moth of the family Crambidae, is widely distributed in 29 countries in Asia, Oceania, Australia and Africa between 48°N and 24°S latitude and 0°E and 172°W longitude, an area with many rice-growing regions with high humidity and temperature (Khan et al., 1988). The rice leaffolder larvae damage crops by cutting and folding the leaves, causing direct yield reduction in rice (Padmavathi et al., 2013). As an important agricultural pest, the rice leaffolder has attracted increasing attention. Many studies have addressed its mechanisms of chemical communication (Liu et al., 2015; Zeng et al., 2015), long-distance migration routes (Wang et al., 2017), environmental adaptation (Bodlah et al., 2019; Chen et al., 2015) and migration–reproduction trade-offs in energy allocation (Wang et al., 2017). Unfortunately, the lack of a high-quality genome assembly hinders in-depth understanding of these biological features.

Lepidoptera and its sister order Trichoptera (caddisflies) are reported to have a female heterogametic system (Carabajal Paladino et al., 2019). Due to the absence of recombination, the sex chromosomes are vulnerable to deleterious mutations. Thus, pseudogenization or gene loss frequently occurs on the sex chromosomes (Ellegren, 2011). It is known that the W chromosome is enriched in repeat sequences (Abe et al., 2005). Hence, assembling a complete W chromosome remains difficult. Although the W chromosomes of several lepidopteran insects have been reported (Fu et al., 2018; Wan et al., 2019; Xiao et al., 2020; Zhang et al., 2019), most are poorly assembled and represent only a fraction of the W chromosome. For example, although several sequencing platforms and a well-established pipeline were used, only a fragment of the entire W chromosome of the codling moth *Cydia pomonella* was assembled (Wan et al., 2019).

Here, we generated a chromosome-level genome assembly of this notorious insect pest by combining several sequencing techniques, namely Illumina short-read sequencing, PacBio long-read sequencing and Hi-C technology. Sex chromosomes were successfully identified. Under enrichment analysis, we found that most W-linked genes were involved in metabolism processes. In addition, a number of gene families of rice leaffolder were observed to be significantly expanded through TREEFAM analysis and CAFE assignment. This genome assembly provides a useful data resource for in-depth analysis of insect migration, chromosome rearrangement and evolution, and pest control.

2 | MATERIALS AND METHODS

2.1 | Insects

Cnaphalocrocis medinalis pupae were collected from a rice field in Jinhua, Zhejiang Province, China, in July 2018, and were maintained at the Zhejiang Academy of Agricultural Sciences. The insect adults were fed with fresh rice leaves and maintained at 26 ± 1°C, under a 14:10-hr (light–dark) photoperiod cycle and 85% ± 5% relative humidity. Five generations were reared and the pupae of the fifth generation were used for sequencing.

2.2 | Genome size and heterozygosity estimation

Genome size of the rice leaffolder was estimated with flow cytometry using a previously reported method (He et al., 2016). A total of 15 female adults were killed, and their heads were used in the experiment. The fruit fly *Drosophila melanogaster* Canton-S strain with a standard genome size of 1C = 176.4 Mb was used as the reference species. Fly or rice leaffolder heads were placed into 500 μl Galbraith buffer (45 mm $MgCl_2$, 30 mm sodium citrate, 20 mm 3-[*N*-morpholino] propane sulfonic acid [MOPS], 1 ml/L Triton X-100), which was adjusted to pH 7.0 using HCl and then filtered through a 0.22-μm nylon filter before use. The tissues were chopped, and the homogenate was filtered through 40U nylon mesh to remove debris. The filtrate was then centrifuged at 2,000 *g* at 4°C for 10 min. Then, the supernatant was removed and 500 μl phosphate buffered saline (PBS) was added to the tubes. The tubes were mixed by shaking before adding 10 μg RNase A into each tube. After 10 min, 25 μg propidium iodide was added to each tube and tubes were covered with tin foil. Then, the solution was placed in the dark for at least 15 min. Flow cytometry was performed on a Partec Cyflow cytometer. All experiments were performed in triplicate.

A genome survey using K-mer analysis was carried out by paired-end sequencing of libraries (350-bp insert size) using the Illumina HiSeq X Ten System. We removed adapters from the reads and discarded low-quality reads including paired-end reads with an N content in a single-end sequencing read exceeding 10% of the read length and reads with fewer than five bases in a single-end sequencing read. Then, 17-mer analysis was carried out with 54.2 Gb clean reads to estimate the genome size and heterozygosity. The 17-mers were extracted base-by-base from each sequence and traversed the entire genome. The distribution of 17-mers obeyed a Poisson distribution and was used to estimate the genome size based on the Lander–Waterman algorithm (Lander & Waterman, 1988). The distribution of 17-mers was also used to estimate heterozygosity. The ratio of the heterozygous peak value to the homozygous peak value was calculated to obtain the heterozygosity.

2.3 | Genome sequencing and assembly

All female pupae were collected and rinsed with precooled 0.9% saline to avoid contamination before being frozen with liquid nitrogen. A total of 10.3 μg genomic DNA was extracted from one female pupa using the sodium dodecyl sulfate (SDS) extraction method (Zhou et al., 1996). Whole-genome shotgun sequencing was performed using PacBio according to the manufacturer's protocol. A Single Molecule Real Time bell DNA library with an insert size of 20 kb was prepared for sequencing. All PacBio reads were

generated by the PacBio Sequel I System with seven cells and were further self-corrected with QUIVER (smrtlink_5.1.0) (https://www.pacb.com/support/software-downloads/). In addition, because of the error rates associated with long reads, we also prepared an Illumina paired-end library (350-bp insert size) and sequenced it with the Illumina HiSeq X Ten System. The clean Illumina short reads were mapped to the assembly profile using BWA version 0.6.2 (Li, 2013), and the long-read assembly was polished using PILON version 1.22 (https://github.com/broadinstitute/pilon/). Genome assembly was conducted using the de Brujin strategy. Specifically, the longest single-pass reads were selected as seed reads and all other reads were mapped to the seed reads using DALIGNER (https://github.com/thegenemyers/DALIGNER). A consensus based on mapped reads was generated with the PBDACGON tool (https://github.com/PacificBiosciences/pbdagcon), and a de novo assembly profile was generated using WTDBG version 1.2.8 (https://github.com/ruanjue/wtdbg-1.2.8). Whole genome sequencing and assembly were performed according to the manufacturer's instructions by Novogene. BUSCO version 3.0.2b (Simao et al., 2015) was used to assess the completeness of the final assembly.

2.4 | Hi-C scaffolding

Following the standard protocol described previously (Belton et al., 2012) with some modifications, Hi-C libraries were constructed using one female pupa as input. The sample was cut into pieces to optimize permeation. After being ground with liquid nitrogen, tissues were cross-linked by incubating in 4% formaldehyde solution at room temperature in a vacuum for 30 min. Glycine (2.5 m) was added to quench the crosslinking reaction for 5 min and then samples were placed on ice for 15 min. The sample was centrifuged at 1,174 *g* at 4°C for 10 min, and the pellet was washed with 500 μl PBS and then centrifuged for 5 min at 1,174 *g*. The pellet was resuspended with 20 μl of lysis buffer (1 m Tris-HCl, pH 8, 1 m NaCl, 10% CA-630, and 13 units protease inhibitor), and the supernatant was then centrifuged at 4,696 *g* at room temperature for 10 min. The pellet was washed twice in 100 μl ice cold 1 × NEB buffer and then centrifuged for 5 min at 4,696 *g*. The nuclei were resuspended in 100 μl NEB buffer, solubilized with dilute SDS, and then incubated at 65°C for 10 min. After quenching the SDS with Triton X-100, the samples were digested with a four-cutter restriction enzyme, *Mbo*I (400 units), at 37°C on a rocking platform overnight.

The DNA ends were marked with biotin-14-dCTP and blunt-end ligation was performed. The proximal chromatin DNA was religated by adding a ligation enzyme. The nuclear complexes were reverse cross-linked by incubating with proteinase K at 65°C. DNA was purified by phenol–chloroform extraction. Biotin was removed from nonligated fragment ends using T4 DNA polymerase. The ends of fragments sheared by sonication (200–600 bp) were repaired with a mixture of T4 DNA polymerase, T4 polynucleotide kinase and Klenow DNA polymerase. Biotin-labelled Hi-C samples were specifically enriched using streptavidin C1 magnetic beads. After adding A-tails to the fragment ends, followed by ligation of the Illumina paired-end sequencing adapters, the Hi-C sequencing library was amplified by PCR (12–14 cycles) and sequenced on the Illumina HiSeq paired-end 150 platform after quality control.

The high-quality sequencing reads were mapped to the draft genome by BWA version 0.6.2 (Li, 2013). Unmapped paired-end reads and multiple mapped reads were filtered by SAMTOOLS version 1.9 (Li et al., 2009). The unique high-quality paired-end reads mapping close to the restriction sites were retained for downstream analysis. LACHESIS (https://github.com/shendurelab/LACHESIS), which is based on the agglomerative hierarchical clustering algorithm (Burton et al., 2013), with default parameters was used to cluster the scaffolds into groups, and the order of the scaffolds was confirmed by the strength of interactions between read pairs. Orientations were assigned to each group. Hi-C sequencing and scaffolding analysis were performed at Novogene. We used the percentage of short reads uniquely mapped to the genome assembly and the coverage rate as criteria, as well as genome completeness generated using BUSCO version 3.0.2b to test the quality of the genome assembly.

2.5 | Synteny analysis

Whole-genome sequence alignments between rice leaffolder and four other lepidopteran insects (the domestic silkworm *Bombyx mori*, the rice stem borer *Chilo suppressalis*, the codling moth *Cydia pomonella* and the beet armyworm *Spodoptera exigua*) were performed using SATSUMA version 3.1.0 (Grabherr et al., 2010) with default parameters. The synteny relationships among chromosomes were displayed using CIRCOS version 0.69–9 (Krzywinski et al., 2009).

2.6 | Identifying sex chromosomes

To identify sex-linked regions in the genome, we performed genome resequencing with five female pupae and five male pupae. High-quality clean reads were obtained with the Illumina paired-end 150 sequencing platform. Genome sequencing was conducted by Annoroad Gene Technology. By comparing the coverage differences between male and female samples as described in Mongue et al. (2017), we identified the Z and W chromosomes of rice leaffolder. Specifically, autosomes have equal coverage in males and females, while there is two-fold greater coverage of the Z chromosome in males. In contrast, the W chromosome shows a female-biased coverage pattern. We mapped the reads to the chromosomal scaffolds with BWA version 0.6.2 (Li, 2013). Then, the R package 'changepoint' version 2.2.2 (https://CRAN.R-project.org/package=changepoint) was used to identify candidate sex chromosomes of *C. medinalis* based on the male:female (M:F) coverage ratio. Chromosomes with a $\log_2$ (M:F read counts) value of 0 were regarded as autosomes, those with a value less than or equal to the threshold |0.25| were considered W linked, and those with a value greater than or equal to |0.25| were considered Z-linked.

2.7 | Genome annotation

To identify repeat elements in the *C. medinalis* genome, a de novo repeat library was constructed using REPEATMODELER version 1.0.11 (Benson, 1999) with integrated results of RECON version 1.08 (Bao & Eddy, 2002), REPEATSCOUT version 1.0.5 (Price et al., 2005) and TRF version 4.09 (Benson, 1999). A de novo repeat library was then generated using REPEATMASKER version 4.0.7 (Tarailo-Graovac & Chen, 2009) and REPBASE (Bao et al., 2015) according to the recommended parameter values. Finally, REPEATMASKER version 4.0.7 was used to identify repetitive elements in the assembled genome. TRNASCAN-SE version 2.0 (Chan & Lowe, 2019) was used to predict transfer RNA (tRNA) genes with eukaryote parameters, and INFERNAL version 1.1.2 software (Nawrocki & Eddy, 2013) was used to search the Rfam database (release 11.0) with an E-value cutoff of 1E-5 to predict noncoding RNAs (ncRNAs).

The Optimized Maker-Based Insect Genome Annotation (OMIGA) pipeline (Liu et al., 2014) was used to annotate the genome of *C. medinalis*. After masking the repeat sequences, ab initio prediction and homology- and transcriptome-based approaches were integrated to predict protein-coding genes. For transcriptome-based prediction, HISAT version 2.1 (Kim et al., 2015) was used to align the transcriptome data (SRR647915, SRR3203088, SRR941774, SRR1793301, SRR1793308, SRR1793305, SRR1793309, SRR1793306, SRR1793307) to the genome, and gene information was predicted using STRINGTIE version 1.3.4c (Pertea et al., 2015). Three ab initio gene prediction programs, AUGUSTUS version 3.3 (Stanke et al., 2006), SNAP version 2006-07-28 (Korf, 2004) and GENEMARK-ET SUITE 4.32 (Lomsadze et al., 2014), were used for de novo gene prediction. To improve the accuracy of the de novo training, intact open reading frames (ORFs) in the transcriptome were identified using TRANSDECODER version 5.0.2 (https://github.com/TransDecoder/TransDecoder/releases), and transcripts with an intact ORF were then used to retrain the ab initio gene prediction program, AUGUSTUS version 3.3 (Stanke et al., 2006). *B.mori*.hmm (https://github.com/KorfLab/SNAP/tree/master/HMM) was selected as the training set when using SNAP version 2006-07-28 (Korf, 2004). For GENEMARK-ET, the whole assembly, which was more than 10 Mb, was used to retrain the software. For the homology-based approaches, the annotated gene sets from all invertebrate species (downloaded from the National Center for Biotechnology Information [NCBI] Refseq database) were aligned to the *C. medinalis* genome using EXONERATE version 2.2.0 (Slater & Birney, 2005) with the protein2genome model (--percent 50 parameter) to refine the BLAST hits for defining exact intron/exon positions. All gene evidence identified from the above three approaches was combined by the MAKER pipeline version 2.31 (Cantarel et al., 2008) into a weighted and nonredundant consensus of gene structures. The default parameters were used for MAKER.

We also searched the SwissProt and NCBI nonredundant databases using BLASTP version 2.8.1 (E-value < 1E-5). For gene ontology (GO) annotations, we aligned all protein-coding sequences to the PANNZER2 database (Törönen et al., 2018) with default settings. For pathway annotation, the Kyoto Encyclopedia of Genes and Genomes (KEGG) database was used to obtain pathway IDs.

To obtain a curated Official Gene Set (OGS), we further shifted out genes whose coding region overlapped repetitive sequences more than 50%. After blasting against to Dfam database (release 3.2) (Hubley et al., 2016) (E-value < 1E-10), genes with each query coverage per matched region more than 50% were removed. Otherwise, we searched the SwissProt annotation and GO annotation using transposon-related keywords and removed corresponding genes.

2.8 | Comparative genomics and phylogenetic reconstruction

The protein sequences of 20 species (Table S1) were used for phylogenomic analysis: namely the rice leaffolder, *C. medinalis*, and 10 other lepidopteran insects (the Japanese oak silk moth, *Antheraea yamamai*, the silkworm, *Bombyx mori*, the codling moth, *Cydia pomonella*, the rice stem borer, *Chilo suppressalis*, the monarch butterfly, *Danaus plexippus*, the tobacco hornworm, *Manduca sexta*, the Asian swallowtail, *Papilio xuthus*, the diamondback moth, *Plutella xylostella*, the tobacco cutworm, *Spodoptera litura*, and the cabbage looper, *Trichoplusia ni*), one Hymenopteran species (the western honey bee, *Apis mellifera*), two Coleoptera insects (the red flour beetle, *Tribolium castaneum*, and the Asian long-horned beetle, *Anoplophora glabripennis*), two Dipteran insects (the common fruit fly, *Drosophila melanogaster*, and the malaria mosquito, *Anopheles gambiae*), two Hemipteran species (the sugarcane aphid, *Melanaphis sacchari*, and the brown planthopper, *Nilaparvata lugens*), the louse, *Pediculus humanus*, and the caddisfly, *Stenopsyche tienmushanensis*. We only kept the longest transcript of each gene for analysis. ORTHOMCL software version 2.0.9 (Li et al., 2003) was used with default settings to identify orthologues and homologues.

To infer the phylogeny of these insects, multiple sequence alignments of single-copy gene families were performed using MAFFT version 7.310 (Katoh & Standley, 2013) with the "-auto" parameter, and trimming was performed by TRIMAL 1.2rev59 (Capella-Gutierrez et al., 2009) with the "-automated1" setting. The alignment results were concatenated to construct a maximum likelihood phylogenetic tree using RAXML version 8.2.10 (Stamatakis, 2014) with the model "LG + F+I + G4." The substitution model was selected using MODELFINDER in IQ-TREE version 1.5.5 (Nguyen et al., 2015). Statistical support was obtained with 1,000 bootstrap replicates. Divergence times between various species were estimated by MCMCtree in PAML 4.9e (Yang, 2007). Four standard divergence time points from the TimeTree database (http://timetree.org/) were used for calibration: (a) *Melanaphis sacchari–Nilaparvata lugens*, 177–401 million years ago (Ma); (b) *Drosophila melanogaster–Anopheles gambiae*, 217–301 Ma; (c) *Stenopsyche tienmushanensis–Plutella xylostella*, 183–278 Ma; and (d) *Tribolium castaneum–Anoplophora glabripennis*, 191–243 Ma. The tree was visualized using FIGTREE version 1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

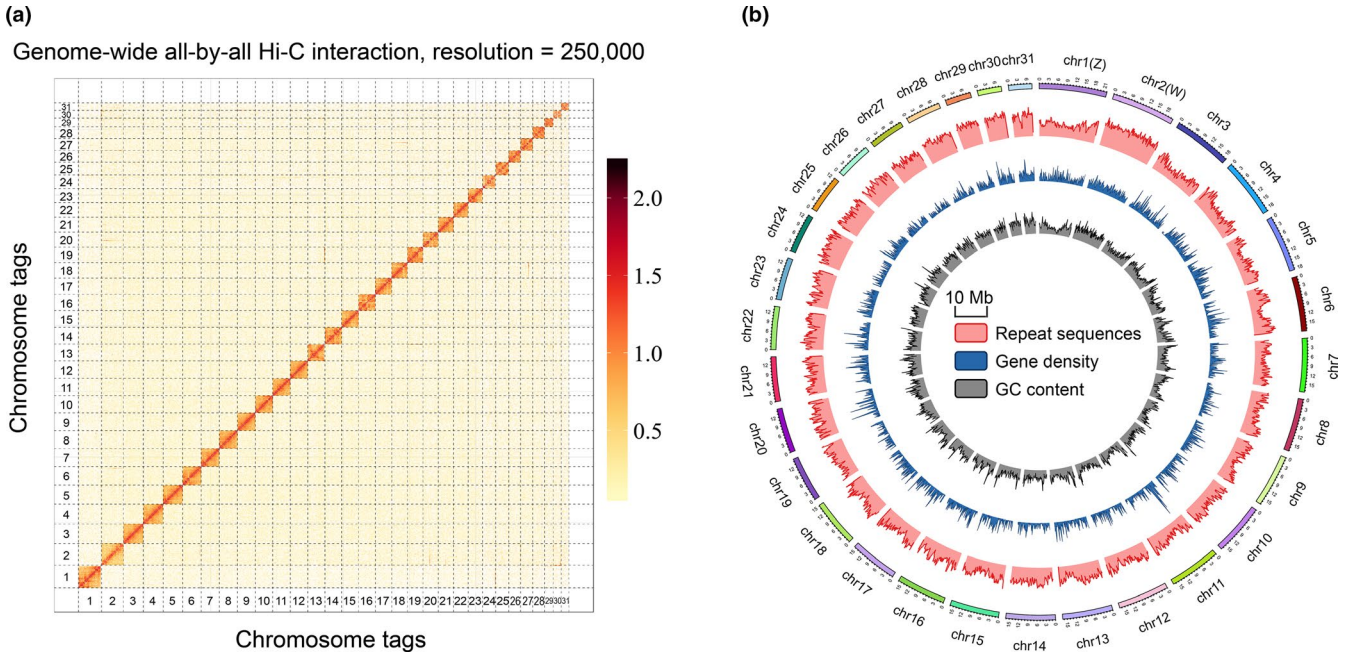

FIGURE 1 Heatmap of genome-wide Hi-C data (resolution = 250,000 bp) and overview of the genomic landscape of rice leaffolder, *Cnaphalocrocis medinalis*. (a) The heatmap shows all-by-all interactions among 31 chromosomes of rice leaffolder. There were strong intrachromosomal interactions (blocks on the diagonal line), while interchromosomal interactions were weaker. The frequency of Hi-C interaction links is represented by the colour, which ranges from white (low) to red (high). (b) Blocks on the outmost circle represent all 31 chromosomes of rice leaffolder. Peak plots from outer to inner circles in red, blue and grey represent repeat sequence coverage, gene density and GC content of each chromosome, respectively (sliding window size = 200 kb)

2.9 | Gene family expansion and contraction

The TreeFam database (Schreiber et al., 2014) was used to analyse the gene number of each gene family in each species. The resulting matrix tables were used as inputs to examine the expansion and contraction of each gene family in CAFE version 4.2.1 (De Bie et al., 2006) with $p < .05$ as the cut-off. The enrichment analysis of gene families was performed using the OMICSHARE tools (http://www.omicshare.com/tools). Additionally, the hypergeometric distribution was used to test the significance of enrichment results and further adjusted by the false discovery rate (FDR).

3 | RESULTS AND DISCUSSION

3.1 | Chromosome-level genome assembly of rice leaffolder

A total of 46.8 Gb of long reads (88.6-fold coverage) was obtained using the PacBio Sequel platform. These PacBio long-reads were self-corrected using QUIVER and 54.3 Gb of Illumina short reads (102.8-fold coverage) was generated for polishing the PacBio reads (Table S2). The self-corrected and polished PacBio reads were used to assemble a draft genome assembly with a total length of 528.3 Mb, consisting of 4,671 contigs with an N50 length of 524.6 kb. The assembled genome size is similar to that obtained by K-mer estimation (513.5 Mb; Figure S1) but larger than that estimated by flow cytometry (434.5 Mb; Figure S2). This inconsistency has been frequently observed in highly heterozygous insects (Elliott & Gregory, 2015). Because the rice leaffolder was estimated to have a high heterozygosity of 2.28% in K-mer analysis, the discrepancies between the genome assembly, K-mer analysis and flow cytometry estimations are possibly due to its high heterozygosity.

Next, we used Hi-C sequencing to orientate and anchor 3,248 scaffolds (89.1% of the whole genome assembly) to 31 chromosomes (Burton et al., 2013; Figure 1a; Tables S3 and S4). The chromosome-level genome assembly was 528.3 Mb with a scaffold N50 of 16.1 Mb. For quality evaluation of the genome assembly, a total of 95.2% of the short reads were uniquely mapped to the genome assembly and the coverage rate was 97.6%, indicating that the assembled genome was of high quality. A BUSCO assessment showed 96.4% of BUSCO genes were successfully detected, of which 92.3% were single copy and 3.0% were duplicated (Table 1). The results of these two evaluations indicated that the genome assembly had a high level of completeness and was suitable for subsequent analysis.

3.2 | Genome annotation

We used REPEATMASKER (Tarailo-Graovac & Chen, 2009) and REPBASE (Bao et al., 2015) to annotate the repeat sequences. In total, 39.5%

of the rice leaffolder genome was annotated as repeat sequences (Figure 1b). Short interspersed nuclear elements (SINEs), long interspersed nuclear elements (LINEs), long terminal repeats (LTRs) and DNA elements accounted for 3.49%, 14.69%, 1.62% and 1.92% of the whole genome, respectively, and 17.83% of repeat sequences were annotated as unclassified (Table S5). A total of 2,194 tRNAs were predicted by TRNASCAN-SE (Chan & Lowe, 2019). Using INFERNAL (Nawrocki & Eddy, 2013), we also identified 27 small nucleolar RNAs (snoRNAs), 80 ribosomal RNAs (rRNAs), 91 small nuclear RNAs (snRNAs), 473 microRNAs (miRNAs), 678 tRNAs and 79 other types of ncRNAs (Table S6).

After masking repeat sequences, a total of 22,433 protein-coding genes were annotated using OMIGA (Liu et al., 2014). We futher removed 7,388 transposable elements associated genes and kept the remaining 15,045 protein-coding genes in OGS for subsequent analysis (Table 2). Of all predicted OGS genes, 94.5% had BLAST hits in the NCBI nonredundant database. Furthermore, 7,104 genes were assigned with GO terms and 5,175 genes were mapped to at least one KEGG pathway.

TABLE 1 BUSCO assessment of the *Cnaphalocrocis medinalis* Hi-C assembly

Category	Number of BUSCOs
Before scaffolding	
C: 95.0% [S: 91.4%, D: 3.6%], F: 1.4%, M: 3.6%	1,658
After scaffolding	
C: 95.3% [S: 92.3%, D: 3.0%], F: 1.1%, M: 3.6%	1,658
Complete BUSCOs (C)	1,580
Complete and single-copy BUSCOs (S)	1,531
Complete and duplicated BUSCOs (D)	49
Fragmented BUSCOs (F)	19
Missing BUSCOs (M)	59

3.3 | Phylogenetic analysis of rice leaffolder and other lepidopteran insects

We collected the protein-coding gene sequences of 20 insects to perform comparative genomics analysis. A total of 22,417 orthologous groups with 582 single-copy orthologous genes were identified using ORTHOMCL; the number of genes assigned to different orthologous groups is displayed in Figure 2 and Table S7. A phylogenetic tree generated using protein-coding sequences of single-copy orthologous genes showed that rice leaffolder and nine other moths were clustered together (Figure 2). The rice leaffolder *Cnaphalocrocis medinalis* and rice stem borer *Chilo suppressalis* formed a sister lineage to Crambidae, while the Japanese oak silk moth *Antheraea yamamai* and the silkworm *Bombyx mori* were in another sister lineage (Kawahara et al., 2019). The split of the Crambidae lineage from other lepidopteran clusters was inferred to be around 74.7 Ma. All 11 lepidopteran insects diverged from the sister lineage caddisfly *Stenopsyche tienmushanensis* about 267.3 Ma ago (Figure 2), which is consistent with a previous report (Luo et al., 2018).

3.4 | Gene family expansion and contraction in rice leaffolder

We used CAFE software to study expansion and contraction of TreeFam gene families (Figure 1). Of the 9,199 gene families in the most recent common ancestor (MRCA) of all 20 species, 1,012 were expanded and 1,525 were contracted in rice leaffolder compared with the common ancestor of rice leaffolder and rice stem borer. In contrast, rice stem borer had only 872 expanded and 811 contracted gene families. Although 563 gene families were contracted in both rice leaffolder and rice stem borer, there were no gene families expanded in both species.

GO enrichment analysis of the 1,012 expanded TreeFam families in rice leaffolder showed that these genes were enriched in processes including organic substance catabolic process (GO: 1901575,

TABLE 2 Summary of the genome assemblies of *Cnaphalocrocis medinalis* and five other lepidopteran insects

Species	*Bombyx mori*	*Cydia pomonella*	*Chilo suppressalis*	*C. medinalis*	*Danaus plexippus*	*Spodoptera exigua*
Assembly size (Mb)	460.3	772.9	825.7	528.3	248.7	446.8
Karyotype	$N = 28$	$N = 28$	$N = 29$	$N = 30$	$N = 30$	$N = 31$
Number of assembled chromosomes	26A + Z	27A + Z+W	28A + Z	29A + Z+W	29A + Z	31A + Z + W
Contig N50 size (Mb)	12.2	0.8	0.3	0.5	0.1	—
Scaffold N50 size (Mb)	16.8	8.9	27.1	16.1	9.2	14.4
Protein-coding genes	16,880	17,184	15,653	15,045	19,762	17,707
Repeats (%)	46.8	42.9	46.4	39.5	—	33.1
GC (%)	38.3	37.4	34.9	38.5	32.1	36.7
Reference	Kawamoto et al., 2019	Wan et al., 2019	Ma et al., 2020	This study	Gu et al., 2019	Zhang et al., 2019

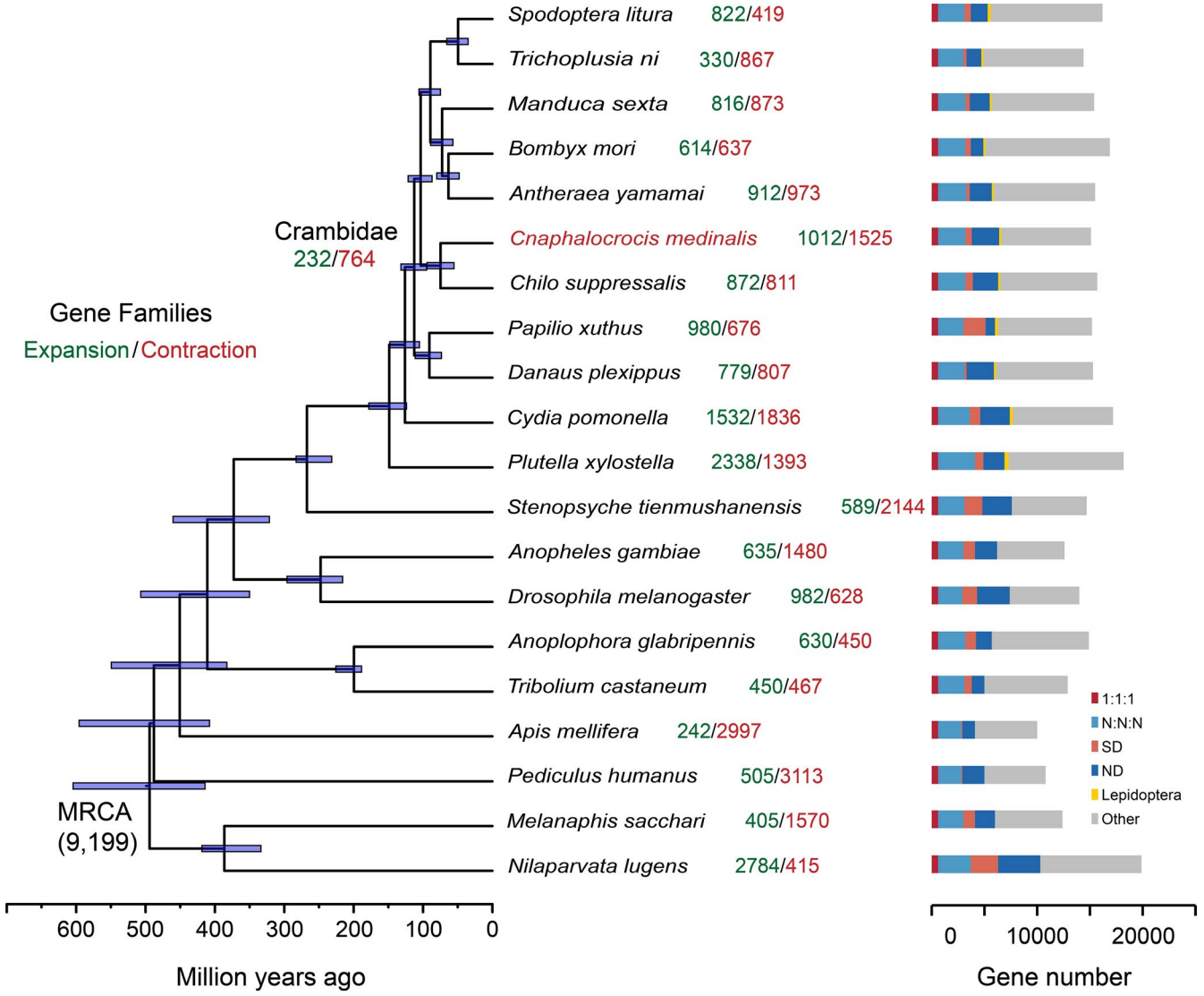

FIGURE 2 Phylogenetic tree and gene orthology. A phylogenetic tree of rice leaffolder *Cnaphalocrocis medinalis* and other insect species was constructed using the maximum likelihood method with concatenated protein sequences of 582 single-copy orthologous genes with 1,000 bootstrap replicates. Divergence times with 95% confidence intervals (blue bars) are displayed at internodes. The numbers of expanded TreeFam gene families (green) and contracted TreeFam gene families (red) are shown to the right of each species branch. MRCA, most recent common ancestor. The coloured bars to the right represent the number of genes classified into six orthology types

225 genes, $p = 1.65E\text{-}08$, FDR-adjusted), cellular defence response (GO: 0006968, 15 genes, $p = 2.82E\text{-}08$, FDR-adjusted), lipid metabolic process (GO: 0006629, 154 genes, $p = 5.45E\text{-}05$, FDR-adjusted) and amino acid transport (GO: 0006865, 23 genes, $p = .00149$, FDR-adjusted; Table S8). KEGG pathway enrichment analysis showed that expanded gene families were significantly enriched in three pathways (ko00100, 15 genes, $p < .01$, FDR-adjusted; ko05217, 13 genes, $p = .01$, FDR-adjusted; ko04142, 36 genes, $p = .02$, FDR-adjusted) and 15 genes were annotated to the steroid biosynthesis pathway in lipid metabolism processes (Table S9).

GO analysis showed that the 1,525 contracted TreeFam gene families were significantly enriched in catabolic processes, metabolic processes, as well as detoxification and biosynthetic processes, such as long-chain fatty-acyl-CoA metabolic process (GO: 0035336, 21 genes, $p = 3.82E\text{-}13$, FDR-adjusted), wax biosynthetic (GO: 0010025, six genes, $p = 3.39E\text{-}05$, FDR-adjusted) and ion transport (GO: 0006811, 75 genes, $p = .00039$, FDR-adjusted) (Table S10). Moreover, these contracted gene families were significantly enriched in 17 KEGG pathways, including cutin, suberine and wax biosynthesis from the Lipid Metabolism Class (ko00073, 15 genes, $p = 1.17E\text{-}15$, FDR-adjusted), the Toll and Imd signalling pathway from the Immune System Class (ko04624, 16 genes, $p = 8.00E\text{-}05$, FDR-adjusted) and fat digestion and absorption from the Digestive System Class (ko04975, eight genes, $p = .00356$, FDR-adjusted) (Table S11). However, further investigations are still needed to determine the functions associated with the genes in these expanded and contracted gene families, such as analysis of their expression patterns and their putative roles in ecological adaptation-associated processes such as migration.

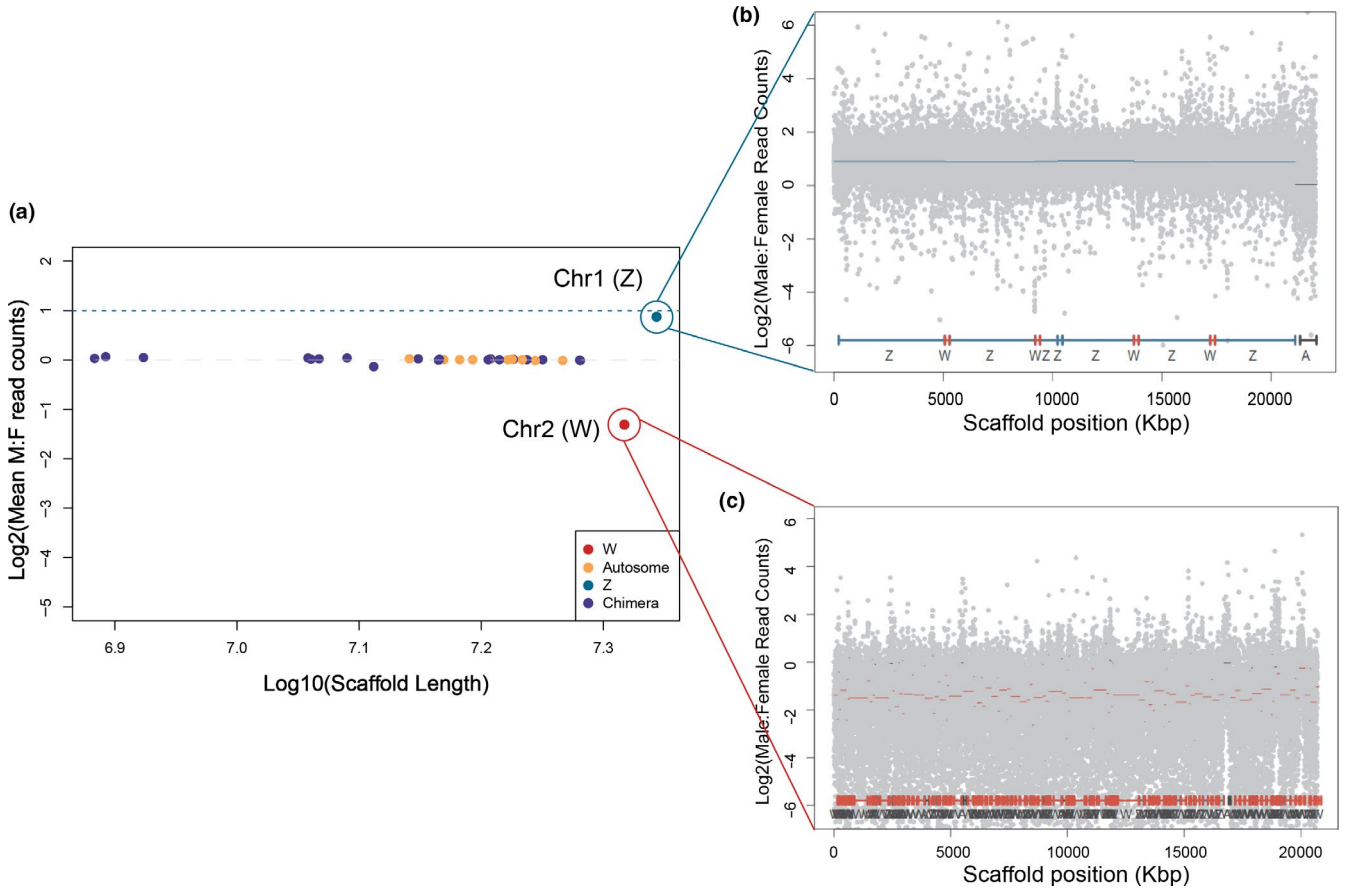

FIGURE 3 Identification of sex chromosomes in the rice leaffolder genome assembly. (a) Male:female (M:F) coverage ratios for each chromosome-scale scaffold, plotted by scaffold length. Each point presents a single chromosome-scale scaffold. The dotted grey line shows the theoretical expectation for autosomes whilst the blue dotted line shows the expectation for the Z chromosome. The orange dots represent autosomal sequences, while the purple dots are chimeric scaffolds. The Z and W chromosomes are circled in green and red, respectively. (b, c) M:F coverage ratios in 500-bp windows across the Chr1 and Chr2 chromosomes. The Z, W and A labels on the *x*-axis indicate the chimeric status (lines coloured with corresponding colours) on Chr1 and Chr2

3.5 | The sex chromosomes of rice leaffolder

We carried out genome resequencing of male pupae and female pupae, producing a total of 51.8 Gb of high-quality clean data with a mean Q30 of 93.3%. The read count coverage ratios were used to identify the Z and W chromosomes (Figure 3a; Mongue et al., 2017). Chr1, the largest chromosome containing 619 scaffolds, had a nearly twofold greater male coverage and was considered to be the Z chromosome (Figure 3b). Chr2, the second largest chromosome containing 171 scaffolds, had a female-biased coverage ratio and was regarded as the W chromosome (Figure 3c). The W chromosome identified in rice leaffolder was estimated to be as long as 20.75 Mb. To the best of our knowledge, this is the most intact W chromosome of lepidopterans known at present (Table S12).

The repeat content of the W chromosome was 11.6 Mb, making up 56% of the entire chromosome. Most of these repeat sequences were LINEs (35%), while LTRs and other DNA elements accounted for 12% and 9.6%, respectively. This was significantly higher than the repeat contents of other chromosomes (0.76%–1.4% for LTRs, 2.28%–3.86% for other DNA elements; Table S13). There were 755 protein-coding genes located on the W chromosome, but 95% of them were uncharacterized proteins. GO enrichment analysis showed that these genes were significantly enriched in three biological processes: DNA integration (GO: 0015074, 17 genes, $p = 5.18E\text{-}06$, FDR-adjusted), DNA recombination (GO: 0006310, 11 genes, $p = 4.01E\text{-}05$, FDR-adjusted) and DNA metabolic process (GO: 0006259, 18 genes, $p = .00401$, FDR-adjusted; Table S14). Moreover, these W-linked protein coding genes were significantly enriched in seven KEGG pathways, namely protein processing in endoplasmic reticulum from the Genetic Information Processing Class (ko04141, two genes, $p = .01619$, FDR-adjusted) and six metabolism pathways related to carbohydrates (ko00052, two genes, $p = .00298$; ko00051, two genes, $p = .00298$; ko00040, two genes, $p = .00298$; all p adjusted by FDR), vitamins (ko00790, two genes, $p = .00298$, FDR-adjusted), lipids (ko00561, two genes, $p = .00298$, FDR-adjusted) and glycan (ko00511, one gene, $p = .03418$, FDR-adjusted; Table S15). Because the information available for W chromosomes and W-linked gene annotations in lepidopterans is still limited, investigations of W-linked genes such as those involved in migration behaviour and their conservation across species will provide new insight into sex-linked phenotypes.

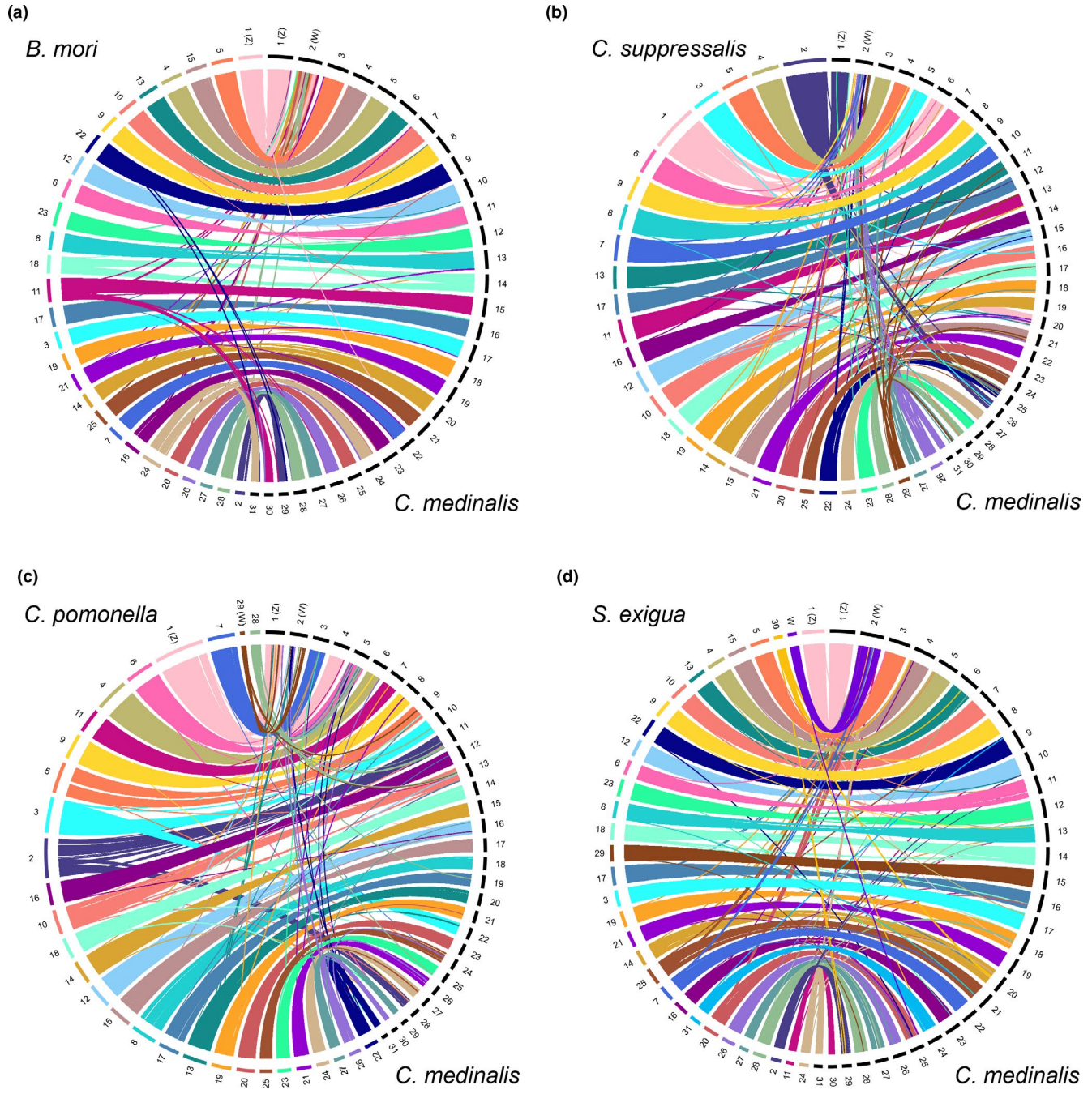

FIGURE 4 Chromosome-level synteny analysis of rice leaffolder (*Cnaphalocrocis medinalis*) and four lepidopteran insects: silkworm (*Bombyx mori*) (a), rice stem borer (*Chilo suppressalis*) (b), codling moth (*Cydia pomonella*) (c), and beet armyworm (*Spodoptera exigua*) (d). The chromosomal arc blocks of *C. medinalis* are shown in black, while those of *Bombyx mori*, *Chilo suppressalis*, *Cydia pomonella* and *Spodoptera exigua* are shown in colour

3.6 | Genome synteny of five lepidopteran insects

We compared the syntenic relationships between rice leaffolder *C. medinalis* and four other lepidopterans, including the silkworm *B. mori*, rice stem borer *Chilo suppressalis*, codling moth *Cydia pomonella* and beet armyworm *Spodoptera exigua*. Although these lepidopteran insects generally shared high chromosomal synteny, we detected a number of fusion and fission events between rice leaffolder and the other lepidopteran insects (Figure 4). High synteny of Z sex chromosomes was observed across the four lepidopteran insects. The rice leaffolder Z chromosome (Chr1) is syntenic to a large portion of the Z chromosome (Chr1) and a small fragment of Chr11 and Chr18 of the silkworm; a large portion of the Z chromosome (Chr2) and a small fragment of Chr2, Chr15 and Chr23 of the rice stem borer; a large portion of the Z chromosome (Chr1) and a small fragment of Chr13, Chr19, Chr26 and the W chromosome (Chr29) of codling

moth; and a large portion of the Z chromosome (Chr1) and a small fragment of Chr19 of beet armyworm. In contrast, many fusion and fission events from both autosomes and sex chromosomes were detected in the W chromosomes of the lepidopteran insects examined. As expected, the rice leaffolder W chromosome shares 99.2% of the syntenic blocks of the beet armyworm W chromosome (Figure 4d), suggesting that these syntenic regions might potentially be ancient W chromosome sequences.

A number of fission and fusion events were also observed in the autosomes. Chr11 in silkworm is syntenic to Chr1, Chr2, Chr15 and Chr30 and a portion of Chr26, Chr27 and Chr29 of rice leaffolder. Small fragments of Chr28, Chr29 and Chr31 and the majority of Chr9 of rice leaffolder share high sequence similarity with Chr22 of silkworm. Silkworm Chr24 is mainly syntenic to Chr31, Chr24 and a portion of the W chromosome of rice leaffolder (Figure 4a). Interestingly, fission events are evident on each chromosome of rice stem borer even though it belongs to the same family, Crambidae, as rice leaffolder. For instance, the entire Chr27 and a part of Chr25 of rice leaffolder are syntenic to Chr24 of rice stem borer, while Chr29 of rice stem borer is syntenic to fragments from 10 chromosomes of rice leaffolder (Figure 4b). The main parts of Chr10 and Chr22 of rice leaffolder are syntenic to Chr3 of codling moth and the main parts of Chr11 and Chr28 of rice leaffolder are syntenic to Chr2 of codling moth (Figure 4c). For beet armyworm, many fusion events covering small regions occurred in all chromosomes (Figure 4d). Beet armyworm has an ancestral karyotype, $n = 31$ (Zhang et al., 2019), and silkworm (Kawamoto et al., 2019) has the same chromosome number ($n = 28$) as codling moth (Wan et al., 2019) whilst rice leaffolder and rice stem borer were reported to have 30 chromosomes (Thakur & Gautam, 2013) and 29 chromosomes (Ma et al., 2020), respectively. The genomic synteny among these five insects illustrates the evolution of karyotype number from 31 to 28 through fusion and fission (Figure 4).

4 | CONCLUSION

We have successfully assembled a high-quality chromosome-level genome assembly of rice leaffolder, which is a helpful resource to study lepidopteran chromosome evolution and the biological characteristics of Crambidae insects, such as migration, cold tolerance, diet breadth and leaf-folding behaviour. We also obtained the most intact W chromosome of lepidopteran insects to date, reflecting a milestone in W chromosome assembly for lepidopteran insects and benefiting studies of sex chromosome evolution and sex determination systems.

ACKNOWLEDGEMENTS

This research was supported by earmarked funds for China Agriculture Research System (CARS-01-36), the State Key Laboratory for Managing Biotic, Chemical Treats to the Quality and Safety of Agro-products (No. 2010DS700124-ZZ2007) and the National Natural Science Foundation of China (31772238). Thanks go to Prof. Jianhua Huang of Zhejiang University for providing the fruit flies for genome size estimation.

CONFLICT OF INTEREST

The authors declare no competing interests.

AUTHOR CONTRIBUTIONS

F.L. conceived and designed the study; F.L., Z.X.L. and H.X.X. coordinated the consortium; H.X.X., Y.J.Y. and X.X.Z. conducted the sampling and sequencing; Z.M.S., X.C., M.Z.L., L.B.G. and X.X.Z annotated the genome; X.X.Z performed the analysis of chromosomal synteny and detection of sex chromosomes, analysis of comparative genomics, and identification of gene families; Y.M. constructed the data website. X.X.Z wrote the draft manuscript; F.L., H.M.X., K.H., X.H.Y., Y.L. and L.X. improved and revised the manuscript.

DATA AVAILABILITY STATEMENT

The BioProject accession no. in the National Center for Biotechnology Information is PRJNA629879. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under accession JABMDO000000000 and the version described in this paper is version JABMDO010000000.The final chromosome assembly and OGSv1 were submitted to InsectBase (http://www.insect-genome.com/Cmed/).

ORCID

Xianxin Zhao https://orcid.org/0000-0002-8704-4000
Xinhai Ye https://orcid.org/0000-0002-0203-0663
Le Xu https://orcid.org/0000-0002-0603-4610
Huamei Xiao https://orcid.org/0000-0003-0165-7410
Fei Li https://orcid.org/0000-0002-8410-5250

REFERENCES

Abe, H., Mita, K., Yasukochi, Y., Oshiki, T., & Shimada, T. (2005). Retrotransposable elements on the W chromosome of the silkworm, *Bombyx mori*. *Cytogenet Genome Res*, *110*(1–4), 144–151. https://doi.org/10.1159/000084946

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, *6*, 11. https://doi.org/10.1186/s13100-015-0041-9

Bao, Z., & Eddy, S. R. (2002). Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Research*, *12*(8), 1269–1276. https://doi.org/10.1101/gr.88502

Belton, J. M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., & Dekker, J. (2012). Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods*, *58*(3), 268–276. https://doi.org/10.1016/j.ymeth.2012.05.001

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research*, *27*(2), 573–580. https://doi.org/10.1093/nar/27.2.573

Bodlah, M. A., Gu, L. L., Wang, G. R., & Liu, X. D. (2019). Rice leaf folder larvae alter their shelter-building behavior and shelter structure in response to heat stress. *Journal of Economic Entomology*, *112*(1), 149–155. https://doi.org/10.1093/jee/toy313

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., & Shendure, J. (2013). Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nature*

Biotechnology, *31*(12), 1119–1125. https://doi.org/10.1038/nbt.2727

Cantarel, B. L., Korf, I., Robb, S. M., Parra, G., Ross, E., Moore, B., Holt, C., Sanchez Alvarado, A., & Yandell, M. (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, *18*(1), 188–196. https://doi.org/10.1101/gr.6743907

Capella-Gutierrez, S., Silla-Martinez, J. M., & Gabaldon, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*(15), 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Carabajal Paladino, L. Z., Provazníková, I., Berger, M., Bass, C., Aratchige, N. S., López, S. N., Marec, F., & Nguyen, P. (2019). Sex chromosome turnover in moths of the diverse superfamily gelechioidea. *Genome Biology Evolution*, *11*(4), 1307–1319. https://doi.org/10.1093/gbe/evz075

Chan, P. P., & Lowe, T. M. (2019). tRNAscan-SE: Searching for tRNA genes in genomic sequences. *Methods in Molecular Biology*, *1962*, 1–14. https://doi.org/10.1007/978-1-4939-9173-0_1

Chen, J., Li, C., & Yang, Z. F. (2015). Identification and expression of two novel cytochrome P450 genes, *CYP6CV1* and *CYP9A38*, in *Cnaphalocrocis medinalis* (Lepidoptera: Pyralidae). *Journal of Insect Science*, *15*, 50. https://doi.org/10.1093/jisesa/ieu174

De Bie, T., Cristianini, N., Demuth, J. P., & Hahn, M. W. (2006). CAFE: A computational tool for the study of gene family evolution. *Bioinformatics*, *22*(10), 1269–1271. https://doi.org/10.1093/bioinformatics/btl097

Ellegren, H. (2011). Sex-chromosome evolution: Recent progress and the influence of male and female heterogamety. *Nature Reviews Genetics*, *12*(3), 157–166. https://doi.org/10.1038/nrg2948

Elliott, T. A., & Gregory, T. R. (2015). What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1678), 20140331. https://doi.org/10.1098/rstb.2014.0331

Fu, Y. U., Yang, Y., Zhang, H., Farley, G., Wang, J., Quarles, K. A., Weng, Z., & Zamore, P. D. (2018). The genome of the Hi5 germ cell line from *Trichoplusia ni*, an agricultural pest and novel model for small RNA biology. *eLife*, *7*, e31628. https://doi.org/10.7554/eLife.31628

Grabherr, M. G., Russell, P., Meyer, M., Mauceli, E., Alfoldi, J., Di Palma, F., & Lindblad-Toh, K. (2010). Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics*, *26*(9), 1145–1151. https://doi.org/10.1093/bioinformatics/btq102

Gu, L., Reilly, P. F., Lewis, J. J., Reed, R. D., Andolfatto, P., & Walters, J. R. (2019). Dichotomy of dosage compensation along the Neo Z chromosome of the monarch butterfly. *Current Biology*, *29*(23), 4071–4077 e4073. https://doi.org/10.1016/j.cub.2019.09.056

He, K., Lin, K., Wang, G., & Li, F. (2016). Genome sizes of nine insect species determined by flow cytometry and k-mer analysis. *Frontiers in Physiology*, *7*, 569. https://doi.org/10.3389/fphys.2016.00569

Hubley, R., Finn, R. D., Clements, J., Eddy, S. R., Jones, T. A., Bao, W., Smit, A. F. A., & Wheeler, T. J. (2016). The Dfam database of repetitive DNA families. *Nucleic Acids Research*, *44*(D1), D81–D89. https://doi.org/10.1093/nar/gkv1272

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kawahara, A. Y., Plotkin, D., Espeland, M., Meusemann, K., Toussaint, E. F. A., Donath, A., Gimnich, F., Frandsen, P. B., Zwick, A., dos Reis, M., Barber, J. R., Peters, R. S., Liu, S., Zhou, X., Mayer, C., Podsiadlowski, L., Storer, C., Yack, J. E., Misof, B., & Breinholt, J. W. (2019). Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(45), 22657–22663. https://doi.org/10.1073/pnas.1907847116

Kawamoto, M., Jouraku, A., Toyoda, A., Yokoi, K., Minakuchi, Y., Katsuma, S., Fujiyama, A., Kiuchi, T., Yamamoto, K., & Shimada, T. (2019). High-quality genome assembly of the silkworm, *Bombyx mori*. *Insect Biochemistry and Molecular Biology*, *107*, 53–62. https://doi.org/10.1016/j.ibmb.2019.02.002

Khan, Z. R., Barrion, A. T., Litsinger, J. A., Castilla, N. P., & Joshi, R. C. (1988). A bibliography of rice leaffolders (Lepidoptera: Pyralidae). *Insect Science and Its Application*, *9*(2), 129–174. https://doi.org/10.1017/S1742758400005919

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360. https://doi.org/10.1038/nmeth.3317

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, *5*, 59. https://doi.org/10.1186/1471-2105-5-59

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., & Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research*, *19*(9), 1639–1645. https://doi.org/10.1101/gr.092759.109

Lander, E. S., & Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, *2*(3), 231–239. https://doi.org/10.1016/0888-7543(88)90007-9

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 [q-bio.GN].

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, L., Stoeckert, C. J. Jr, & Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, *13*(9), 2178–2189. https://doi.org/10.1101/gr.1224503

Liu, J., Xiao, H., Huang, S., & Li, F. (2014). OMIGA: Optimized Maker-Based Insect Genome Annotation. *Molecular Genetics and Genomics*, *289*(4), 567–573. https://doi.org/10.1007/s00438-014-0831-7

Liu, S., Rao, X. J., Li, M. Y., Feng, M. F., He, M. Z., & Li, S. G. (2015). Glutathione S-transferase genes in the rice leaffolder, *Cnaphalocrocis medinalis* (Lepidoptera: Pyralidae): Identification and expression profiles. *Archives of Insect Biochemistry and Physiology*, *90*(1), 1–13. https://doi.org/10.1002/arch.21240

Lomsadze, A., Burns, P. D., & Borodovsky, M. (2014). Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Research*, *42*(15), e119. https://doi.org/10.1093/nar/gku557

Luo, S., Tang, M., Frandsen, P. B., Stewart, R. J., & Zhou, X. (2018). The genome of an underwater architect, the caddisfly *Stenopsyche tienmushanensis* Hwang (Insecta: Trichoptera). *Gigascience*, *7*(12), giy143. https://doi.org/10.1093/gigascience/giy143

Ma, W., Zhao, X., Yin, C., Jiang, F., Du, X., Chen, T., Zhang, Q., Qiu, L., Xu, H., Joe Hull, J., Li, G., Sung, W. K., Li, F., & Lin, Y. (2020). A chromosome-level genome assembly reveals the genetic basis of cold tolerance in a notorious rice insect pest, *Chilo suppressalis*. *Molecular Ecology Resources*, *20*(1), 268–282. https://doi.org/10.1111/1755-0998

Mongue, A. J., Nguyen, P., Volenikova, A., & Walters, J. R. (2017). Neo-sex chromosomes in the monarch butterfly, *Danaus plexippus*. *G3 (Bethesda)*, *7*(10), 3281–3294. https://doi.org/10.1534/g3.117.300187

Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, *29*(22), 2933–2935. https://doi.org/10.1093/bioinformatics/btt509

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Padmavathi, C., Katti, G., Padmakumari, A. P., Voleti, S. R., & Subba Rao, L. V. (2013). The effect of leaffolder *Cnaphalocrocis medinalis* (Guenée)

[Lepidoptera: Pyralidae] injury on the plant physiology and yield loss in rice. *Journal of Applied Entomology*, *137*(4), 249–256. https://doi.org/10.1111/j.1439-0418.2012.01741.x

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*(3), 290–295. https://doi.org/10.1038/nbt.3122

Price, A. L., Jones, N. C., & Pevzner, P. A. (2005). *De novo* identification of repeat families in large genomes. *Bioinformatics*, *21*(Suppl 1), i351–i358. https://doi.org/10.1093/bioinformatics/bti1018

Schreiber, F., Patricio, M., Muffato, M., Pignatelli, M., & Bateman, A. (2014). TreeFam v9: A new website, more species and orthology-on-the-fly. *Nucleic Acids Research*, *42*(D1), D922–D925. https://doi.org/10.1093/nar/gkt1055

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Slater, G. S., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, *6*, 31. https://doi.org/10.1186/1471-2105-6-31

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Research*, *34*(suppl_2), W435–439. https://doi.org/10.1093/nar/gkl200

Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*, Chapter 4, Unit 4.10. https://doi.org/10.1002/0471250953.bi0410s25

Thakur, R., & Gautam, D. C. (2013). Chromosome studies on four species of moths. *Cytologia*, *78*(3), 327–331. https://doi.org/10.1508/cytologia.78.327

Törönen, P., Medlar, A., & Holm, L. (2018). PANNZER2: A rapid functional annotation web server. *Nucleic Acids Research*, *46*(W1), W84–W88. https://doi.org/10.1093/nar/gky350

Wan, F., Yin, C., Tang, R., Chen, M., Wu, Q., Huang, C., Qian, W., Rota-Stabelli, O., Yang, N., Wang, S., Wang, G., Zhang, G., Guo, J., Gu, L. A., Chen, L., Xing, L., Xi, Y. U., Liu, F., Lin, K., … Li, F. (2019). A chromosome-level genome assembly of *Cydia pomonella* provides insights into chemical ecology and insecticide resistance. *Nature Communications*, *10*(1), 4237. https://doi.org/10.1038/s41467-019-12175-9

Wang, F., Yang, F., Lu, M., Luo, S.-Y., Zhai, B.-P., Lim, K.-S., McInerney, C. E., & Hu, G. (2017). Determining the migration duration of rice leaf folder (*Cnaphalocrocis medinalis* (Guenée)) moths using a trajectory analytical approach. *Scientific Reports*, *7*, 39853. https://doi.org/10.1038/srep39853

Xiao, H., Ye, X., Xu, H., Mei, Y., Yang, Y., Chen, X., Yang, Y., Liu, T., Yu, Y., Yang, W., Lu, Z., & Li, F. (2020). The genetic adaptations of fall armyworm *Spodoptera frugiperda* facilitated its rapid global dispersal and invasion. *Molecular Ecology Resources*, *20*(4), 1050–1068. https://doi.org/10.1111/1755-0998.13182

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591. https://doi.org/10.1093/molbev/msm088

Zeng, F.-F., Zhao, Z.-F., Yan, M.-J., Zhou, W., Zhang, Z., Zhang, A., Lu, Z.-X., & Wang, M.-Q. (2015). Identification and comparative expression profiles of chemoreception genes revealed from major chemoreception organs of the rice leaf folder, *Cnaphalocrocis medinalis* (Lepidoptera: Pyralidae). *PLoS One*, *10*(12), e0144267. https://doi.org/10.1371/journal.pone.0144267

Zhang, F., Zhang, J., Yang, Y., & Wu, Y. (2019). A chromosome-level genome assembly for the beet armyworm (*Spodoptera exigua*) using PacBio and Hi-C sequencing. *bioRxiv*. https://doi.org/10.1101/2019.12.26.889121

Zhou, J., Bruns, M. A., & Tiedje, J. M. (1996). DNA recovery from soils of diverse composition. *Applied and Environmental Microbiology*, *62*(2), 316–322. https://doi.org/10.1002/bit.260490302

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Zhao X, Xu H, He K, et al. A chromosome-level genome assembly of rice leaffolder, *Cnaphalocrocis medinalis*. *Mol Ecol Resour*. 2020;00:1–12. https://doi.org/10.1111/1755-0998.13274