Chromosomal-Level Genome Assembly of a True Bug, *Aspongopus chinensis* Dallas, 1851 (Hemiptera: Dinidoridae)

Tao Jiang [1,2], Zhiyong Yin[1,2], Renlian Cai[1], Hengmei Yu[2], Qin Lu[2], Shuai Zhao[1], Ying Tian[1], Yufang Yan[2], Jianjun Guo[1,2,*], and Xiangsheng Chen[1,2,*]

[1]Institute of Entomology, Guizhou University, Guiyang, China
[2]Guizhou Provincial Key Laboratory for Agricultural Pest Management of the Mountainous Region, Guiyang, China

*Corresponding authors: E-mails: jjguo@gzu.edu.cn; chenxs3218@163.com.

Abstract

The true bug, *Aspongopus chinensis* Dallas, 1851 (Hemiptera: Dinidoridae), is a fascinating insect with prolonged diapause and medicinal properties but also a notorious pest. However, because of the lack of genomic resources, an in-depth understanding of its biological characteristics is lacking. Here, we report the first genome assembly of *A. chinensis* anchored to 10 pseudochromosomes, which was achieved by combining PacBio long reads and Hi-C sequencing data. This chromosome-level genome assembly was 1.55 Gb in size with a scaffold N50 of 156 Mb. The benchmarking universal single-copy ortholog (BUSCO) analysis of the assembly captured 96.6% of the BUSCO genes. A total of 686,888,052 bp of repeat sequences, 18,511 protein-coding genes, and 1,749 noncoding RNAs were annotated. By comparing the *A. chinensis* genome with that of 8 homologous insects and 2 model organisms, 213 rapidly evolving gene families were identified, including 83 expanded and 130 contracted gene families. The functional enrichment of Gene Ontology and KEGG pathways showed that the significantly expanded gene families were primarily involved in metabolism, immunity, detoxification, and DNA/RNA replication associated with stress responses. The data reported here shed light on the ecological adaptation of *A. chinensis* and further expanded our understanding of true bug evolution in general.

Key words: PacBio sequence, whole-genome sequence, genome annotation, gene family evolution, Hi-C.

Significance

True bugs (Hemiptera: Heteroptera) are vital insect pests that affect public health and agronomy, yet, the whole-genome assemblies of only five species are available. Here, we assembled a complete draft genome of the true bug *Aspongopus chinensis* using the PacBio sequencing technology and used Hi-C sequences to assist chromosomal assembly. The whole-genome assembly produced in this study could be a useful genomic resource for understanding the evolutionary biology and biological characteristics of true bugs in general and of *A. chinensis* in particular. This assembly can be used to develop effective pest control strategies.

Introduction

Hemiptera is an insect order that is distributed across the world and contains >8,000 nonholometabolous species (Capinera 2008). Hemiptera is classified into five suborders: Sternorrhyncha (aphids, scale insects, psyllids, and whiteflies), Fulgoromorpha (planthoppers), Cicadomoirpha (leafhoppers, cicadas, and spittlebugs), Coleorrhyncha (moss bugs), and Heteroptera (true bugs) (Wang et al. 2019). True bugs are sap-sucking insects that consume food from sources ranging from animals to plants (Sparks et al. 2020). *Aspongopus chinensis* Dallas, 1851 (supplementary fig. S1), a true bug species mainly distributed in Eastern South Asia, is a polyphagous insect that has a prolonged diapause (Luo et al. 2012; Li et

al. 2020). These two characteristics contribute to its adaptation efficiency to harsh environments. Although it is a severe threat to pumpkin and watermelon as a pest, *A. chinensis* has medicative and economic value; it is used in treating nephritis and gastric as well as menstrual pain (Wu et al. 2021).

In the present study, we report the first chromosome-scale genome of a true bug from the Dinidoridae family, *A. chinensis*, which was sequenced using the Pacific Biosciences (PacBio) SMRT sequencing technology. In the genome, we annotated the repetitive sequences, protein-coding genes, and noncoding RNAs. In addition, we analyzed the gene family evolution of hemipteran clades based on the genome and annotated the function of the rapidly evolving gene families of *A. chinensis*. The high-quality *A. chinensis* genome reported in this study represents an essential genetic resource for improving researchers' understanding of the biological characteristics of the species and for conducting Hemipteran evolutionary studies.

Results and Discussion

Genome Assembly and Completeness

In total, 140 Gb of filtered Illumina clean data were used to perform k-mer basis analysis. When $k = 21$, the estimated genome size of predicted genome was 1.36 Gb (heterozygous ratio = 1.71) (supplementary fig. S2). First, we obtained 134 Gb of PacBio clean data with 99.2-fold coverage and assembled these into 15,818 contigs (contig N50 = 602 kb). Then, using the Hi-C technology, the final assembly (size = 1.55 Gb; contig N50 = 1.498 Mb; scaffold N50 = 156 Mb) was produced, which was 13.97% larger than the estimated genome size (i.e., 1.36 Gb). A benchmarking universal single-copy ortholog (BUSCO) evaluation, which demonstrates genome completeness using the genome mode "-m genome," found 96.6% orthologs based on insecta_odb10 (94.9% complete and single-copy, 1.7% duplicated, 0.6% fragmented, and 2.8% missing BUSCO genes) (supplementary tables S1–S4). Subsequently, we reconstructed Hi-C libraries to produce 107.2 Gb of high-quality clean data (Q30 ≥ 91%), 73% of which were unique paired-end reads (supplementary table S5). After de novo assembly, 15,818 contigs were anchored to 1,901 scaffolds. Finally, we obtained 10 pseudo-chromosomes sized 42–215 Mb (fig. 1a and supplementary fig. S3) that accounted for 91.48% of the former genome assembly (1.55 Gb).

Gene Annotation

We annotated 44.3% of the whole genome as repeat sequences, including DNA elements (6.73%), long interspersed nuclear elements (22.15%), long terminal repeats (0.03%), short interspersed nuclear elements (SINEs) (27.82%), and unclassified repeat sequences (0.77%). With a total length of 431,415,958 bp, SINEs were the richest type of repeat sequence. SINEs play crucial roles in genome evolution and modulation of gene expression (Kanhayuwa and Coutts 2016).

We predicted 1,749 noncoding RNAs, including 418 ribosomal RNAs (rRNAs), 656 transfer RNAs (tRNAs), 17 small nuclear RNAs (snRNAs), and 40 microRNAs (miRNAs) (supplementary table S6). After aligning the predicted genes with data obtained from various protein databases (i.e., the non-redundant database, SwissProt, KEGG, TrEMBL, InterPro, and Gene Ontology), we found 18,511 predicted genes. Each gene contained an average of 2.69 exons; a total of 50,654 exons and 31,863 introns were found. The mean lengths of mRNA, exons, and introns were 5,727.22, 413.24, and 2,585.13 bp, respectively.

Gene Families and Phylogenetic Analyses

We identified 165,029 genes from 10 insect species in 18,640 orthogroups. Overall, 6,500 species-specific orthogroups (16.9% in total) were found. The gene set cluster analysis of four heteropteran species (*Cimex lectularius*, *A. chinensis*, *Apolygus lucorum*, and *Halyomorpha halys*) showed that they shared 4,122 clusters. Furthermore, *A. chinensis* contained 582 unique clusters (3,008 proteins), accounting for 4.94% of all clusters of the four species; these could be attributable to novel proteins and unique horizontal gene transfer (fig. 1b). Based on 1,269 single-copy orthogroups, a phylogenetic tree was constructed that included divergence time and orthologous gene families (fig. 1c). The phylogenetic tree and divergence time were consistent with hemipteran taxonomy and results from a previous study (Liu et al. 2021). As shown in the tree, two suborders of Hemiptera, namely Fulgoromorpha and Heteroptera, diverged from a recent common ancestor at 225–242 Ma and from Aphidomorpha approximately at 265–280 Ma. Moreover, the true bugs *H. halys* and *A. chinensis* appear to have diverged at <50 Ma, indicating that they have extraordinarily close consanguinity.

A total of 1,755 expanded and 1,113 contracted genes were identified in 83 significantly expanded and 130 contracted gene families, respectively. The functional annotation of the 83 significantly expanded gene families indicated their association with metabolism, environmental adaptation, digestion, transport, and immune and detoxification systems (fig. 2a). Metabolic pathways were significantly enriched in *A. chinensis*; these are likely associated with overwintering behavior and metamorphosis (Klowden 2013). In addition, dynamic changes in nutrients are generally used by insects to regulate the diapause period (Hahn and Denlinger 2011).

Further analyses showed that the expanded gene families enriched in GO (fig. 2b) and KEGG (fig. 2c) primarily functioned in metamorphosis, immunity, DNA/RNA replication, detoxification, and metabolism, which may explain the mechanism of the broad phytophagy of *A. chinensis* as well as its adaptation to adverse environmental conditions. One
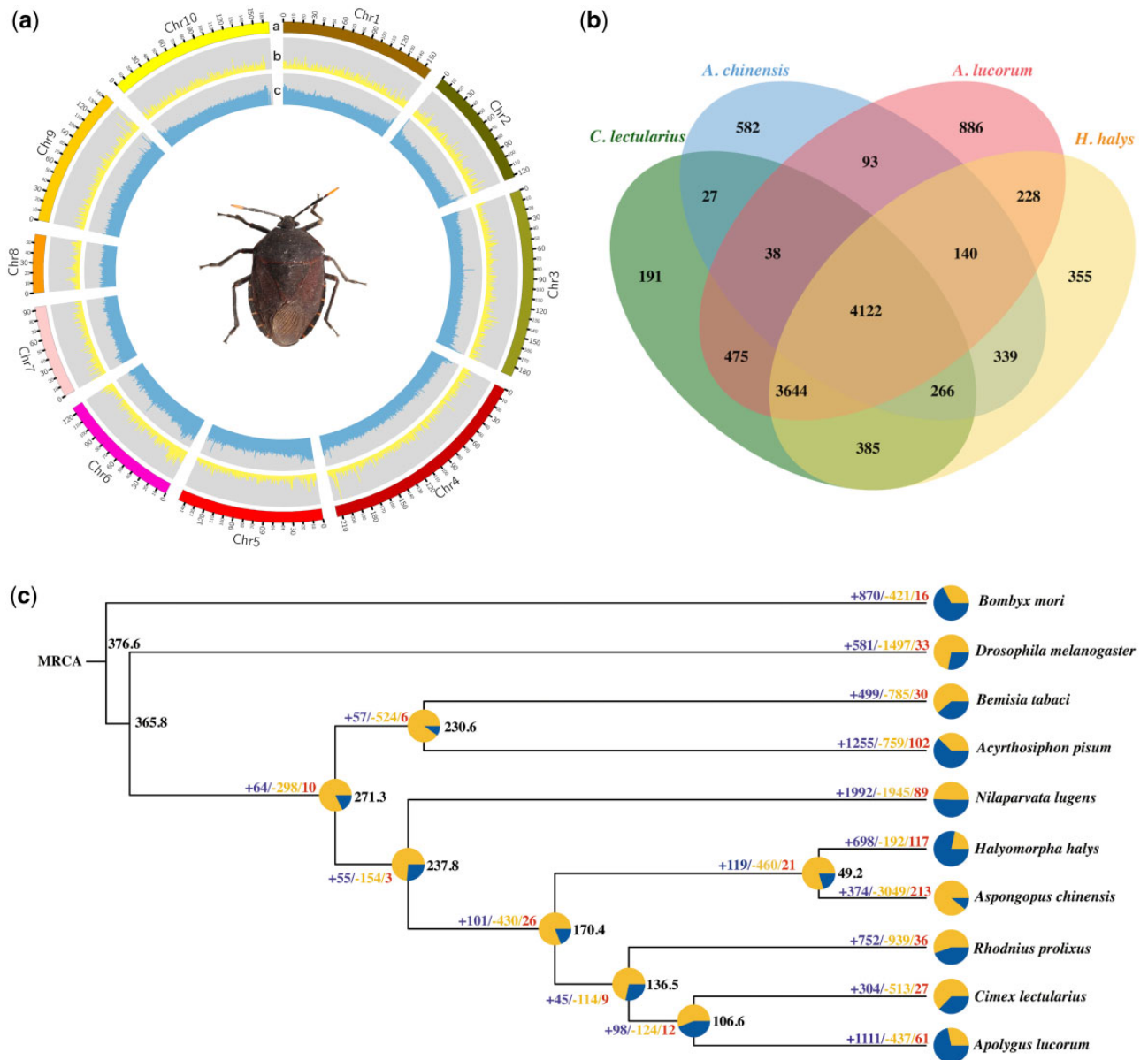
Fig. 1.—Genomic and phylogenetic analysis of *A. chinensis*. (*a*) The genome landscape of *A. chinensis*. Tracks a, b, and c represent chromosome ideograms, gene density, and guanine–cytosine content (%; sliding window size: 500 kb), respectively. (*b*) Venn diagram of orthologous clusters showing the intersection between *A. chinensis* and three homologous insects. (*c*) Phylogeny of 10 selected insect species. The blue, yellow, and red node values represent expanded, contracted, and rapidly evolving families, respectively. Black numbers on the node show divergence time (Ma).

specifically identified KEGG pathway, retinol metabolism, is associated with color vision and therefore essential for habitat exploration and foraging (Lebhardt and Desplan 2017). Furthermore, metabolic pathways (including carbohydrate, lipid, cofactor and vitamin, ascorbate and aldarate, and energy metabolisms), are significantly enriched in diapausic insects compared with non-diapausic *A. chinensis* groups (Qi et al. 2015; Chen et al. 2017; Wu et al. 2021). In the KEGG clustered pathways, genes involved in the HIF-1 signaling pathway are significantly expressed in diapause-destined pupal brains (Lin et al. 2016). Overall, these analyses provide further insights

into the genomics of *A. chinensis*, particularly in terms of its feeding traits and environmental adaptation.

Materials and Methods

Sample Collection and Sequencing

Aspongopus chinensis samples were collected from Kaili City, Guizhou Province, China (26°39′16.65″N, 107°47′18.73″E) and subsequently reared for at least three generations under controlled conditions (25 °C, 80% relative humidity, and a 14:10-h light:dark photoperiod) at the Institute of
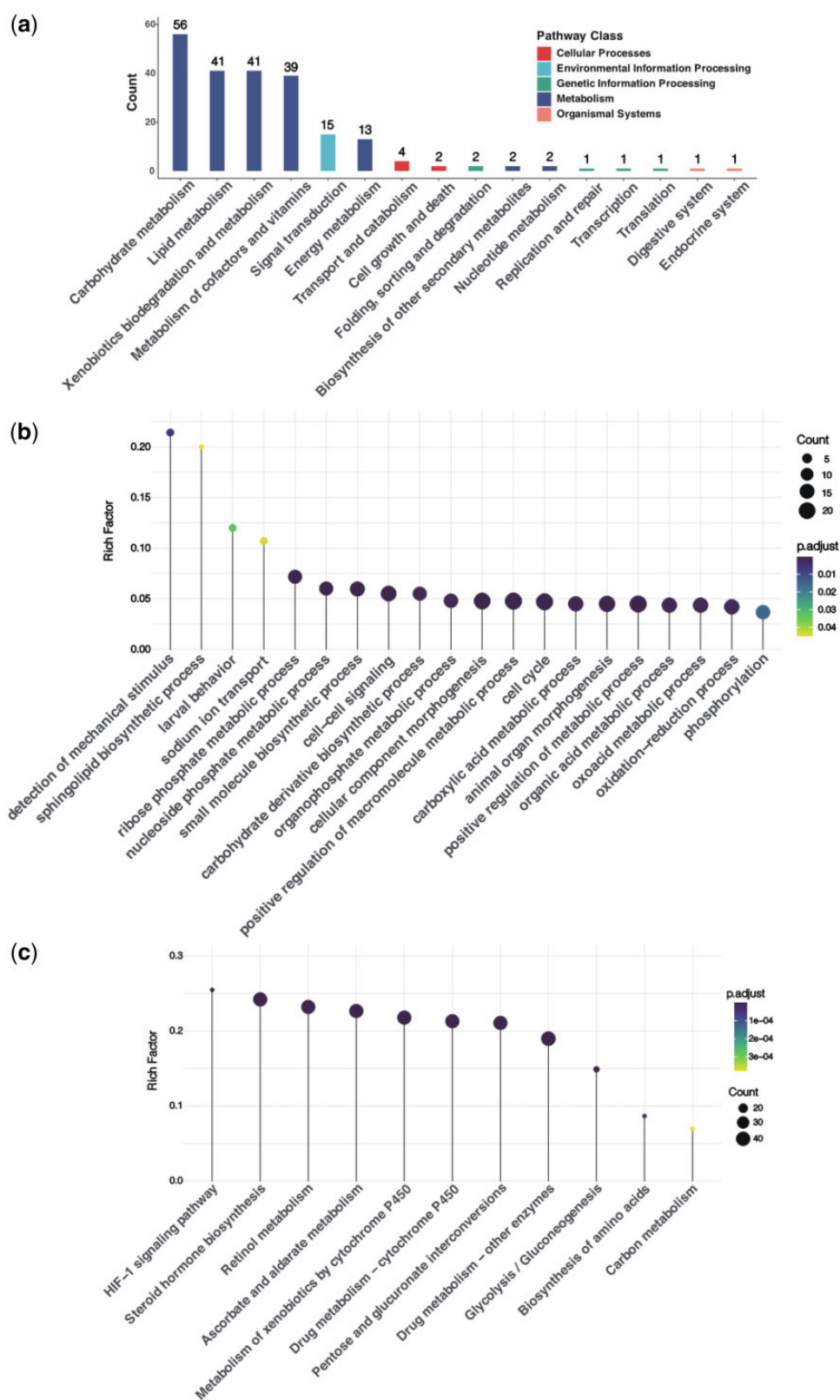
Fig. 2.—Functional annotation of significantly expanded gene families. (a) Eighty-three significantly expanded gene families were annotated in the pathways with the number of genes and categorized into five different pathway classes. (b) Top 20 function enrichment gene ontology terms. (c) KEGG pathways for significantly expanded gene families.

Entomology, Guizhou University. To survey the genome size and heterozygous ratio, we constructed a 350-bp insert-size DNA library for Illumina sequencing on the HiSeq NovaSeq 6000 platform, a 20-kb insert-size library sequencing using the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, CA, USA), and a 150 paired-end RNA library using TRIzol Reagent. We filtered the raw reads with short lengths (<50 bp) and low-quality values (<0.8) and then removed the adapter sequence consistencies.

Chromosomal Genome Assembly

The Illumina short-read sequencing data were used to count k-mers using Jellyfish v2.1.3 and GenomeScope v1.0 (Andrew 2010; Marçais and Kingsford 2011). We used CANU v1.8 (Koren et al. 2017) to assemble the genome with PacBio clean data and corrected errors using Arrow v1.5.5 in one round (Sparks et al. 2020). We polished the resultant assembly with Pilon v1.2 (Walker et al. 2014) and cleaned redundancy based on Hi-C sequences using trimDup (Rabbit Genome Assembler; https://github.com/gigascience/rabbit-genome-assembler, last accessed October 13, 2021). The genome assembly was aligned to remove bacterial contamination using BLAT v3.2.20.

After filtering low-quality sequences ($Q \leq 15$), adapter sequences, and short sequences (length ≤ 30 bp) using HiC-Pro v2.8.1 (Servant et al. 2015), the Hi-C data were mapped to the draft assembly genome using bowtie v2-2.2.5 (Langmead and Salzberg 2012) to obtain unique mapped (Kent 2002) paired-end reads. We then generated chromosomal scaffolds using Juicer v1.5 and 3D de novo assembly pipelines to assemble Hi-C clean data (Durand et al. 2016). The completeness of the genome assembly was evaluated using BUSCO v5.1.2 (Simão et al. 2015).

Genome Annotation

RepeatMasker v4.0.7 (Tarailo-Graovac and Chen 2009) was used to detect transposable elements against a de novo repeat library built using RepeatModeler v2.0.1. Tandem Repeats Finder (Benson 1999) was used to mark tandem repeat sequences in the *A. chinensis* genome. rRNAs were predicted using BLASTN (Chen et al. 2015), whereas tRNAs were annotated using tRNAscan-SE v1.3.1(Lowe and Eddy 1997). Furthermore, we used INFERNAL v1.1 (Griffiths-Jones et al. 2005) based on Rfam v14.0 to annotate miRNAs and snRNAs.

The annotation of *A. chinensis* RNA-seq data and the protein sequences of the homologous species *Oncopeltus fasciatus*, *Rhodnius prolixus*, *H. halys*, *Acyrthosiphon pisum*, *C. lectularius*, and *Nilaparvata lugens* (supplementary table S7) were used to train a novel gene model via AUGUSTUS v3.2.3 (Stanke et al. 2006) and SNAP v2013-02-16 (Johnson et al. 2008). A nonredundant annotated gene set was obtained by integrating and filtering the results using Maker v2.31.10 (Holt and Yandell 2011). A circos plot was drawn on 10 chromosomes using CIRCOS v0.69-9 (Krzywinski et al. 2009). Gene

functional annotation was performed by mapping the predicted gene set to specific protein databases (i.e., SwissProt, TrEMBL, KEGG, and InterPro) using Blast v2.2.31 (Kent 2002).

Gene Family and Phylogenetic Analyses

One dipteran species, *Drosophila melanogaster*, one lepidopteran species, *Bombyx mori*, and seven hemipteran species (Fulgoromorpha: *N. lugens*, *A. pisum*, and *Bemisia tabaci*; Heteroptera: *R. prolixus*, *C. lectularius*, *A. lucorum*, and *H. halys*) were selected for gene family analyses. We clustered the orthologous gene families using OrthoFinder v2.3.8 with the parameters "-S diamond -M msa -T fastree" (Emms and Kelly 2019). We then aligned the protein sequences of the 10 species with MUSCLE v3.8.1551 using its default options (Katoh and Standley 2013). Next, we concatenated the orthologous genes to a single alignment using RAxML v1.5.5 (Jarvis et al. 2014) (-m PROTGAMMAJTT -# 1000). The MCMCTree v4.9 of PAML (Yang 2007) was used to perform the divergence time analysis and calibration time was based on three nodes: *D. melanogaster* with *A. lucorum* (320–390 Ma) (Misof et al. 2014), *A. pisum* with *A. lucorum* (260–280 Ma) (International Aphid Genomics Consortium 2010; Liu et al. 2021), and *H. halys* with *R. prolixus* (144–338 Ma) (www.timetree.org/). The expansion/contraction of gene families was estimated using CAFÉ v4.2.1 (Han et al. 2013), with the single birth–death parameter lambda and a significance level of 0.01. In addition, we used OthoVenn2 (Xu et al. 2019) to cluster the gene sets of *A. chinensis*, *C. lectularius* (Rosenfeld et al. 2016), *A. lucorum*, and *H. halys* with E-values of $1e^{-5}$ and an inflation value of 1.5.

The function of rapidly expanded gene family was annotated using eggNOG-mapper v2 (Cantalapiedra et al. 2021). We performed GO and KEGG enrichment analyses using the R package clusterProfiler v3.18.1 (Yu et al. 2012) and Enrichplot v1.10.2 (P-value $= 0.05$; q-value $= 0.05$).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

Data Availability

The genome assembly has been submitted to NCBI under the BioProject PRJNA729875. Genome annotation and manually annotated families are available at figshare (https://figshare.

com/s/7bc3ec0b7fff879ecf6e, last accessed October 13, 2021). The comparative analysis scripts run in this study are available in the following repository: https://github.com/JT2303/Achi_genome/tree/master (last accessed October 13, 2021).

Literature Cited

Andrew S. 2010. A quality control tool for high throughput sequence data. Fast QC. 390(391). Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27(2):573–580.

Capinera JL. 2008. Harlequin Bug, Murgantia histrionica (Hahn) (Hemiptera: Pentatomidae). In: Capinera JL, editor. Encyclopedia of Entomology. Netherlands: Springer. p. 1766–1768.

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol. doi: 10.1093/molbev/msab293.

Chen Y, Ye W, Zhang Y, Xu Y. 2015. High speed BLASTN: an accelerated MegaBLAST search tool. Nucleic Acids Res. 43(16):7762–7768.

Chen Y, et al. 2017. Transcriptome sequencing reveals potential mechanisms of diapause preparation in bivoltine silkworm Bombyx mori (Lepidoptera: Bombycidae). Comp Biochem Physiol Part D Genomics Proteomics. 24:68–78.

Durand NC, et al. 2016. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 3(1):95–98.

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference or comparative genomics. Genome Biol. 20(1):238.

Griffiths-Jones S, et al. 2005. Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res. 33(Database issue):D121–124.

Hahn DA, Denlinger DL. 2011. Energetics of insect diapause. Annu Rev Entomol. 56:103–121.

Han MV, Thomas GW, Lugo-Martinez J, Hahn MW. 2013. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol Biol Evol. 30(8):1987–1997.

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 12:491.

International Aphid Genomics Consortium. 2010. Genome sequence of the pea aphid Acyrthosiphon pisum. PLoS Biol. 8:e1000313.

Jarvis ED, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science. 346(6215):1320–1331.

Johnson AD, et al. 2008. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics. 24(24):2938–2939.

Kanhayuwa L, Coutts RH. 2016. Short interspersed nuclear element (SINE) sequences in the genome of the human pathogenic fungus Aspergillus fumigatus Af293. PLoS One. 11(10):e0163215.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30(4):772–780.

Kent WJ. 2002. BLAT—the BLAST-like alignment tool. Genome Res. 12(4):656–664.

Klowden MJ. 2013. Physiological systems in insects. London: Academic Press. p. 324–345.

Koren S, et al. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27(5):722–736.

Krzywinski M, et al. 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19(9):1639–1645.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods. 9(4):357–359.

Lebhardt F, Desplan C. 2017. Retinal perception and ecological significance of color vision in insects. Curr Opin Insect Sci. 24:75–83.

Li S, et al. 2020. Advances in studies on chemical constituents, pharmacological effects and clinical application of Aspongopus chinensis. China J Chinese Materia Medica. 45:303–311.

Lin XW, Tang L, Yang J, Xu WH. 2016. HIF-1 regulates insect lifespan extension by inhibiting c-Myc-TFAM signaling and mitochondrial biogenesis. Biochim Biophys Acta. 1863(11):2594–2603.

Liu Y, et al. 2021. Apolygus lucorum genome provides insights into omnivorousness and mesophyll feeding. Mol Ecol Resour. 21(1):287–300.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25(5):955–964.

Luo XH, et al. 2012. The biosynthetic products of Chinese insect medicine, Aspongopus chinensis. Fitoterapia. 83(4):754–758.

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 27(6):764–770.

Misof B, et al. 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science. 346(6210):763–767.

Rosenfeld JA, et al. 2016. Genome assembly and geospatial phylogenomics of the bed bug Cimex lectularius. Nat Commun. 7:10164.

Qi X, et al. 2015. De novo transcriptome sequencing and analysis of Coccinella septempunctata L. in non-diapause, diapause and diapause-terminated states to identify diapause-associated genes. BMC Genomics. 16(1):1–16.

Servant N, et al. 2015. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. 16:259–269.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31(19):3210–3212.

Sparks ME, et al. 2020. Brown marmorated stink bug, Halyomorpha halys (Stål), genome: putative underpinnings of polyphagy, insecticide resistance potential and biology of a top worldwide pest. BMC Genomics. 21(1):227.

Stanke M, et al. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 34(Web Server issue):W435–W439.

Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics. Chapter 4:Unit 4.10. Unit 4.10.

Walker BJ, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 9(11):e112963.

Wang YH, et al. 2019. When did the ancestor of true bugs become stinky? Disentangling the phylogenomics of Hemiptera-Heteroptera. Cladistics. 35(1):42–66.

Wu Y, et al. 2021. Differential metabolism of juvenile hormone III between diapause and non-diapause of Aspongopus chinensis Dallas (Hemiptera: Dinidoridae) revealed by transcriptome sequencing. J Asia Pac Entomol. 24(2):199–204.

Xu L, et al. 2019. OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. Nucleic Acids Res. 47(W1):W52–W58.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24(8):1586–1591.

Yu G, Wang LG, Han Y, He QY. 2012. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 16(5):284–287.

Associate editor: Dennis Lavrov