Chromosome-level genome assembly of the sweet potato weevil, Cylas formicarius (Fabricius) (Coleoptera: Brentidae) and functional characteristics of CforOBP4-6

Jinfeng Hua[1], Lei Zhang[1], Xiaofeng Dong[1], Yonghua Han [1], Xiaowan Gou[1], Jianying Sun[1], Yujie Fu[1], Huifeng Li[2], Yongmei Huang[2], Yanqing Li[2], Tianyuan Chen[2], Daifu Ma[3], and Zongyun Li[1]

[1]School of Life Sciences Jiangsu Normal University
[2]Maize Research Institute, Guangxi Academy of Agricultural Sciences
[3]Xuzhou Academy of Agricultural Sciences/Sweet Potato Research Institute, CAAS

June 25, 2021

Abstract

Cylas formicarius is one of the most important pests of sweet potato worldwide, causing considerable ecological and economic damage. To improve the effect of comprehensive management and understanding of genetic mechanisms, the genetic functions of C. formicarius have been the subject of intensive study. Using Illumina and PacBio sequencing, we obtained a chromosome-level genome assembly of adult weevils from lines inbred for 15 generations. The high-quality assembly obtained had a size of 338.84 Mb, with contig and scaffold N50 values of 14.97 Mb and 34.23 Mb, respectively. In total, 157.51 Mb of repeat sequences and 11,907 protein-coding genes were predicted. A total of 337.06 Mb of genomic sequences was located on the 11 chromosomes, and the sequence length that could be used to determine the sequence and direction accounted for 99.03% of the total length of the associated chromosome. Comparative genomic analysis showed that C. formicarius was sister to Dendroctonus ponderosae, and C. formicarius diverged from D. ponderosae approximately 138.89 million years ago (Mya). Many important gene families that were expanded in the C. formicarius genome were involved in the chemosensory system. In an in-depth study, the binding assay results indicated that CforOBP4-6 had strong binding affinities for sex pheromones and other ligands. Overall, the high-quality C. formicarius genome provides a valuable resource to reveal the molecular ecological basis, genetic mechanism and evolutionary process of major agricultural pests, deepen the understanding of environmental adaptability and apparent plasticity, and provide new ideas and new technologies for ecologically sustainable pest control.

Chromosome-level genome assemblyof the sweet potato weevil,*Cylas formicarius* (Fabricius) (Coleoptera: Brentidae), and functional characteristics of CforOBP4-6

Jinfeng Hua[1,2]| Lei Zhang[1]| Xiaofeng Dong[1]| Yonghua Han[1]| Xiaowan Gou[1]|Jianying Sun[1]| Yujie Fu[1]| Huifeng Li[2]|Yongmei Huang[2]| Yanqing Li[2]| Tianyuan Chen[2*]| Daifu Ma[3*]| Zongyun Li[1*]

[1]Institute of Integrative Plant Biology, Jiangsu Key Laboratory of Phylogenomics & Comparative Genomics, School of Life Sciences, Jiangsu Normal University, Jiangsu, China

[2]Sweet Potato Laboratory, Maize Research Institute, Guangxi Academy of Agricultural Sciences, Guangxi, China

[3]Xuzhou Academy of Agricultural Sciences/Sweet Potato Research Institute, CAAS , Jiangsu, China.

* Correspondence

Zongyun Li

Institute of Integrative Plant Biology, Jiangsu Key Laboratory of Phylogenomics & Comparative Genomics, School of Life Sciences, Jiangsu Normal University, No.101 Shanghai Road, Tongshan New District, Xuzhou City, Jiangsu Province, China, E-mail: *zongyunli@jsnu.edu.cn*

Daifu Ma

Xuzhou Academy of Agricultural Sciences/Sweet Potato Research Institute, CAAS, North of the high-speed railway station, Xuhai Road, Xuzhou Economic and Technological Development Zone, Xuzhou City, Jiangsu Province, China, E-mail:*daifuma@163.com*

Tianyuan Chen

Sweet Potato Laboratory, Maize Research Institute, Guangxi Academy of Agricultural Sciences, No.174 Daxue Road, Nanning City, Guangxi Province, China, E-mail: *ctyuan@gxaas.net*

Abstract

Cylas formicarius is one of the most important pests of sweet potato worldwide, causing considerable ecological and economic damage. To improve the effect of comprehensive management and understanding of genetic mechanisms, the genetic functions of *C .formicarius* have been the subject of intensive study. Using Illumina and PacBio sequencing, we obtained a chromosome-level genome assembly of adult weevils from lines inbred for 15 generations. The high-quality assembly obtained had a size of 338.84 Mb, with contig and scaffold N50 values of 14.97 Mb and 34.23 Mb, respectively. In total, 157.51 Mb of repeat sequences and 11,907 protein-coding genes were predicted. A total of 337.06 Mb of genomic sequences was located on the 11 chromosomes, and the sequence length that could be used to determine the sequence and direction accounted for 99.03% of the total length of the associated chromosome. Comparative genomic analysis showed that *C . formicarius* was sister to*Dendroctonus ponderosae* , and *C . formicarius*diverged from *D . ponderosae* approximately 138.89 million years ago (Mya). Many important gene families that were expanded in the*C . formicarius* genome were involved in the chemosensory system. In an in-depth study, the binding assay results indicated that CforOBP4-6 had strong binding affinities for sex pheromones and other ligands. Overall, the high-quality *C. formicarius* genome provides a valuable resource to reveal the molecular ecological basis, genetic mechanism and evolutionary process of major agricultural pests, deepen the understanding of environmental adaptability and apparent plasticity, and provide new ideas and new technologies for ecologically sustainable pest control.

Keywords

Cylas formicarius , PacBio, Hi-C, chromosome-level genome, functional annotation, odorant binding proteins

1 |Introduction

Brentidae, members of which are also called primitive weevils**,**includes over 11,000 described extant species, including many of the world's insect pest species (https://bugguide.net/node/view/15740) (Bouchard et al., 2011; Gunter, Oberprieler, & Cameron, 2016; Schon & Skuhrovec, 2016). Weevils are responsible for substantial damage to food and cash crops, causing severe reduction in crop yields and considerable economic loss (Christiaens et al., 2016; Hardee, Jones, & Adams, 1999; Industry, 2016; Kyereko, Hongbo, Amoanimaa-Dede, Meiwei, & Yeboah, 2019). Sweet potato weevil (SPW),*Cylas formicarius* (Fabricius) (Coleoptera: Brentidae), is a major destructive pest that causes drastic yield decline, resulting in a loss of millions of dollars annually. Although olfaction-based strategies have been used to prevent and control infestations of the sweet potato weevil as part of integrated pest management (IPM) programmes (Coffelt, Vick, Sower, & Mcclellan, 1978; Heath et al., 1986), they exhibit a unique ability to damage sweet potato throughout tropical and subtropical regions of the world (Kyereko *et al.* , 2019). To provide a wealth of information to improve the effect of comprehensive management and understand the molecular ecology and evolution of this species, the genetic functions of *C. formicarius* have been the subject of intensive study.

Sweet potato (*Ipomoea batatas* [L.] Lam), the seventh most important crop in the world and the fourth most significant crop in China (Food and Agriculture Organization of the United Nations), is an important source of calories, proteins, vitamins and minerals for humans (J. Yang et al., 2017). Sweet potato has immense potential to play a major role in human nutrition, food security, and poverty alleviation in developing countries (Bovell-Benjamin, 2007). *C . formicarius* is the major destructive pest of sweet potato throughout Africa, Asia, the Pacific islands, the Caribbean, USA, Venezuela and Guyana (Hiroyoshi, Kohama, & Reddy, 2016; Kyereko*et al.* , 2019) and has been found in higher-latitude areas as well (Korada & Mukherjee, 2012). Although*C. formicarius* prefers sweet potato, more than 30 species of*Ipomoea* and other genera have been recorded as its host plants (McConnell & Hossner, 1991; Sutherland, 1986). In southern China (Jiangsu, Zhejiang, Jiangxi, Hunan, Fujian, Guizhou, Sichuan, Yunnan, Guangxi, Guangdong, Hainan and Taiwan), *C . formicarius* can produce several generations per year and overwinter in storage or in open fields (Ma, Wang, Li, Gao, & Chen, 2016) After originating in the Indian subcontinent approximately 80-100 million years ago (Mya) (Wolfe, 1991), *C . formicarius*first became associated with sweet potato, which originated in or near northwestern South America, at the beginning of the sixteenth century (Austin, 1988). *C . formicarius* was first described in 1792-1794 by Fabricius from Trenquebar (India), and it was first reported as an economic pest in 1857 (Cockerham, Deen, Christian, & Newsom, 1954). Over the course of 150 years of research, many studies on *C . formicarius* management and control have been carried out, including studies on agricultural measures, chemical and biological control, host plant resistance, insect sterilization techniques and IPM.

Furthermore, the sex pheromone of *C. formicarius* , (Z)-3-dodecen-1-yl (E)-2-butenoate, was first extracted in 1978 (Coffelt*et al.* , 1978), and the bioactivity of the synthetic compound was tested in both laboratory and field experiments in 1986 (Heath *et al.* , 1986). Olfaction-based approaches, using synthetic sex pheromones to monitor and interfere with the pests' ability to find suitable mates, have been used successfully in "push-pull" control strategies (Hlerema, Laurie, & Eiasu, 2017). However, because*C. formicarius* populations have overlapping generations and because these insects are active throughout the year, have unknown larval feeding behaviour, fly short distances, and are mainly nocturnal as adults, there are no effective control strategies for application in the production of sweet potato (Kyereko *et al.* , 2019). Generally, in quarantine areas worldwide, *C. formicarius*causes extensive loss of sweet potato (Kyereko *et al.* , 2019; Ondiaka, Maniania, Nyamasyo, & Nderitu, 2008). Thus, there is an urgent need to develop alternative pest control methods. So far, the development of different but related gene expression patterns has been reported based on transcriptome analysis (Bin, Qu, Pu, Wu, & Lin, 2017; Ma *et al.* , 2016); however, genomic research on *C . formicarius* , including on the mechanism of environmental adaptation and the molecular mechanism of olfactory recognition, has been very limited.

In the present study, we report a chromosome-level genome assembly of*C .formicarius* using Illumina paired-end (PE) sequencing, Pacific Biosciences (PacBio) long reads and High-throughput chromosome conformation capture technology (Hi-C) chromatin interaction maps. The high-quality genome sequence will provide a strong foundation for the biological study of *C . formicarius* , which will advance the knowledge of the mechanisms of molecular evolution, host-plant specialization, ecological adaptation and innovative pest control.

2 | Materials and Methods

2.1 | Insects

C .formicarius(NCBI Taxonomy ID: 2611,543) (Figure 1) was collected from Nanning, Guangxi, China, followed by 15 generations of single-pair mating with fresh sweet potato roots at a temperature of 27±1 °C under 60±5% relative humidity (RH) and a 16:8 h light-dark (L:D) photoperiod at the School of Life Sciences, Jiangsu Normal University.

2.2 |Genome sequencing

High-molecular-weight genomic DNA was isolated from approximately 400 male and female weevils using a TreliefTM Animal Genomic DNA Kit (TsingKe, China), and the DNA quality and quantity was assessed using a NanoDrop spectrophotometer (Thermo Fisher Scientific, USA) and a Qubit(r) 3.0 Fluorometer

(Invitrogen, USA). The extracted DNA was then used to construct Illumina libraries and PacBio RSII libraries. PE genomic libraries with an insertion length of 270 bp were constructed and sequenced on an Illumina HiSeq X Ten platform (Illumina, USA) according to the Illumina TruSeq Nano DNA Library Prep Kit; a total of 27.64 Gb of raw data of PE sequences (2 x 150 bp) for *C. formicarius* were obtained. For long-read sequencing, single molecule real-time (SMRT) cell 20-kb DNA libraries were constructed on a PacBio Sequel sequencer (Pacific Biosciences, Menlo Park, CA, USA) according to the standard PacBio protocol; one movie of the SMRT cells was acquired at Biomarker Technologies Corporation (Beijing, China). The original data were subjected to strict quality control before assembly. For Illumina data, we used the Trimmomatic program v0.33 (Trimmomatic, RRID:SCR-011848) to remove adaptor sequences and trim low-quality reads (Bolger, Lohse, & Usadel, 2014).

2.3 | Genome evaluation, assembly and correction

The sequence data (27.64 Gb of clean reads) from 270-bp PE libraries were employed in the analysis of the k-mer (k=19) depth frequency distribution map. The size of the *C. formicarius* genome was estimated as follows: G = K-num/K-depth (where G represents the genome size, K-num is the total number of 19-mers, and K-depth is the average k-mer depth) (Marcais & Kingsford, 2011). The 19-mer peak was at a depth of 61X. The estimated genome size was used to obtain the subsequent genome assembly results.

The PacBio long-read data were assembled using an overlap-layout-consensus method (Staden, 1980). First, the longer reads were selected and corrected, and these were then used to obtain a draft assembly. Second, the draft assembly was polished. PacBio long reads were assembled and corrected using the Canu program (v 1.7) (Koren et al., 2017). To obtain their maximum supported range, error-corrected reads were obtained by trimming unsupported bases and hairpin adapters with default parameters, and then, the draft assembly was generated. Preliminary assembly was conducted using Falcon v1.2.4 (Koren *et al.* , 2017). To further improve the quality of the reference assembly, PacBio data were assembled into contigs with the Wtdbg program (Ruan & Li, 2020). We obtained a 338.84 Mb genome assembly and contig N50 of 14.97 Mb. To further improve the accuracy of the assembly and evaluate the genomic integrity, four rounds of consensus correction were performed using Illumina reads mapped with Burrows-Wheeler Aligner (BWA v0.7.10-r789) (H. Li & Durbin, 2009), Pilon (Pilon, RRID:SCR-014731) (Walker et al., 2014) software, CEGMA v2.5 (http:// korflab.ucdavis.edu/ Datasets) (Parra, Bradnam, & Korf, 2007) and BUSCO v2.0 (http://busco.ezlab.org) (Simao, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015).

2.4 |Hi-C library construction and chromosome assembly based on Hi-C data

To generate the linear chromosome-level assembly of the *C. formicarius* genome, we constructed the Hi-C fragment library ranging from 300 to 700 bp by following previously described protocols (Nagano et al., 2015; Rao et al., 2014). Insect tissue was fixed with 2% (vol/vol) formaldehyde in PBS, and the DNA was digested with HindIII. The sticky ends were biotinylated and proximity-ligated to form chimaeric junctions, which were enriched and further sheared into 300-700 bp fragments by sonication. These chimaeric fragments were sequenced using a PE strategy on the Illumina HiSeq X Ten platform at Biomarker Technologies Corporation (Beijing, China). To obtain clean data, adapter sequences and low-quality PE reads were removed using FastQC software (Andrews, 2014). Then, the clean reads were mapped using BWA v0.7.10-r789 (H. Li & Durbin, 2009; Walker *et al.* , 2014) (Table S4). Only uniquely aligned PE reads were considered for subsequent analysis. Identification and filtering of the invalid read pairs, sorting and quality assessment were performed using HiC-Pro (v2.11.1) (Servant et al., 2015). By using Lachesis software, the verified data were used to group, cluster, sort and orient the contigs into chromosome-level sequences (Burton et al., 2013). (Table S5, Figure 1).

2.5 |Genome sequence annotation

Repeat sequences are less well conserved among species and play an important role in genome evolution (Treangen & Salzberg, 2012). Based on the assembled genome, we used the RepeatScout v1.05 (Price, Jones, & Pevzner, 2005), PILER-DF v2.4 (Edgar & Myers, 2005), LTR-FINDER v1.05 (Z. Xu & Wang, 2007) and MITE-Hunter v1.0.0 (Han & Wessler, 2010) software packages to construct a de novo repeat

library with default parameters. First, we used TRF (v.4.09), RepeatMasker (v. 3.3.0; RepeatMasker, RRID:SCR 012954) and Repeat Protein Mask (v. 3.3.0) to detect repeat sequences and classify different types of repetitive sequences by aligning genome sequences to the Repbase library (v. 17.01) (Bao, Kojima, & Kohany, 2015). Next, we conducted a RepeatModeler analysis on the de novo library and used RepeatMasker (v4.0.6) (Tarailo-Graovac & Chen, 2009). Then, PASTEClassifier (Wicker et al., 2007) was used to classify the repeat libraries, and the Repbase database (Jurka et al., 2005) was used to merge the libraries. Finally, we used RepeatMasker v4.0.6 to identify the repeat regions by aligning sequences against the Repbase and de novo repeat libraries.

After masking the repeat sequence, de novo-based, homologue-based and RNA sequence-based gene methods were employed for gene prediction in the*C . formicarius* genome assembly. We first used the software programs Genscan v3.1 (Burge & Karlin, 1997), Augustus v3.0 (Augustus, RRID:SCR_008417) (Stanke, Steinkamp, Waack, & Morgenstern, 2004), GlimmerHMM v3.0.4 (Majoros, Pertea, & Salzberg, 2004), SNAP v2006-07-28 (Korf, 2004) and GeneID v1.4 (Blanco, Parra, & Guigo, 2007) for de novo prediction. The protein sequences of the coleopteran insects *C .formicarius* , *Anoplophora glabripennis* ,*Dendroctonus ponderosae* , *Oryctes borbonicus* ,*Tribolium castaneum* and *Drosophila melanogaster*were downloaded from the National Center for Biotechnology Information (NCBI) database (https://www.ncbi.nlm.nih.gov/genome/), and homology-based prediction was performed using GeMoMa v1.3.1 (Keilwagen et al., 2016) as a reference.

In the process of RNA sequence-based gene prediction, eleven RNA samples were extracted from larvae and 3-day-old adult tissues (antennae, heads, legs, thoraxes and abdomens) of males and females using a V Total RNA Isolation System kit (Promega Madison, WI, USA). The RNA-seq reads of *C . formicarius* from Illumina sequences were assembled with the reference genome using Hisat V2.0.4 (Kim, Langmead, & Salzberg, 2015) and Stringtie V1.2.3 (Pertea, Kim, Pertea, Leek, & Salzberg, 2016). After filtering, TransDecoder v2.0 (http:// trans decod er.github.io), GenemarkS-T v5.1(Tang, Lomsadze, & Borodovsky, 2015) and Program to Assemble Spliced Alignments (PASA, RRID:SCR_014656) (Haas et al., 2003) were used for RNA-seq-based gene prediction. Finally, the results of gene annotation from the three approaches were integrated by EVidenceModeler (EVM) v1.1.148 (EVM, RRID:SCR_014659) (Haas et al., 2008). To avoid the loss of some reliable genes in EVM integration, the genes lost in EVM integration were added based on the predicted results of the de novo-based, homologue-based and RNA sequence-based methods and modified with PASA v2.0.2 to obtain the whole-genome assembly.

According to the structural characteristics of different non-coding RNAs, different strategies (rRNA, microRNA and tRNA prediction) were adopted to predict the *C . formicarius* genome. The rRNAs were predicted using RNAmmer v1.2 software by aligning the *C .formicarius* genome to the Rfam database (release 13.0) (Kalvari et al., 2018). The tRNAs were identified using tRNAscan-SE v1.3.1 (tRNAscan-SE, RRID: SCR_010835) software with default parameters for eukaryotes (Lowe & Eddy, 1997). Based on the miRBase database (Griffiths-Jones, Grocock, van Dongen, Bateman, & Enright, 2006), Infinal 1.1(Nawrocki & Eddy, 2013) was used to predict microRNAs. The results of non-coding RNA annotation are presented in Table S14.

For the annotation of pseudogenes, we searched for sequences homologous to the known protein-coding genes in the *C. formicarius* genome using Genblasta V1.0.4 (She, Chu, Wang, Pei, & Chen, 2009). The premature termination codons or frameshift mutations located in the above sequences were identified and pseudogenes were obtained using Genewise V2.4.1 (RRID:SCR 015054) (Birney, Clamp, & Durbin, 2004). A total of 503 pseudogenes were annotated in the genome of *C. formicarius* .

Gene functional annotation was performed by alignment to the Eukaryotic Orthologous Groups of proteins (KOG) database (Tatusov et al., 2001), nucleotide collection (nr/nt) (Marchler-Bauer et al., 2011), TrEMBL database (Boeckmann et al., 2003), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Figure S4) (Kanehisa & Goto, 2000) and Swiss-Prot protein knowledgebase (http:// www.expasy.org/sprot/ and http://www.ebi.ac.uk/ swissprot/) using Basic Local Alignment Search Tool (BLAST) v2.2.31 (Altschul, Gish, Miller, Myers, & Lipman, 1990) and KAAS v2.1(Marcais & Kingsford, 2011). Furthermore, InterProScan v5.8-49.0 (RRID:SCR 005829) (Jones et al., 2014) was used to annotate conserved functional motifs

and protein domains, and the functional annotations were aligned to the following databases: PROSITE (RRID: SCR003457) (Bairoch, 1991), PRINTS (RRID: SCR 003412) (Attwood & Beck, 1994), SUPERFAMILY (Gough & Chothia, 2002), PANTHER (RRID: SCR 004869) (Mi et al., 2005), TIGRFAMs (Haft, Selengut, & White, 2003), SMRT 4.0 (RRID:SCR 005026) (Letunic et al., 2004), PIRSF (C. H. Wu et al., 2004), ProDom (RRID: SCR 006969) (Bru et al., 2005), Pfam (RRID: SCR 004726) (Finn et al., 2014), HAMAP (Lima et al., 2009) and CATH-Gene3D (Lees et al., 2012). Finally, 1386 motifs and 24493 protein domains were annotated in the genome of*C. formicarius* .

2.6 | Comparative genomic analysis

We used the whole-genome sequence of*C. formicarius* and 15 published whole-genome sequences, namely, those of*T.castaneum* (Richards et al., 2008), *Onthophagustaurus* (Choi et al., 2010), *D . ponderosae* (Keeling et al., 2013), *A. glabripennis* (McKenna et al., 2016),*O. borbonicus* (Meyer et al., 2016),*Agrilusplanipennis* (Duan et al., 2019), *Bombyx mori* (Xia et al., 2004),*Apis mellifera* (Sequencing Consortium, 2006),*Locusta migratoria* (Wang et al., 2014),*D.melanogaster* (Gelbart, 1992), *Acyrthosiphon pisum*(International Aphid Genomics, 2010),*Pediculus humanus*(Pittendrigh et al., 2006), *Cimex lectularius* (Rosenfeld et al., 2016),*Zootermopsisnevadensis* (Terrapon et al., 2014), and*Caenorhabditiselegans* (Consortium, 1998), to predict orthologs and infer a phylogenetic tree. To identify the conserved orthologues, we aligned all the protein sequences translated from the longest transcripts of each gene in pairwise using BLASTP (E-value cut-off of 1e-5). The BLASTP results were used to cluster gene families and 1:1 orthologous gene sets in OrthoMCL (L. Li, Stoeckert, & Roos, 2003). Multiple alignments were performed for each orthologue group of the coding sequences of the single-copy families using MAFFT (Katoh, Misawa, Kuma, & Miyata, 2002). Using the orthologous single-copy genes of the 16 species, we connected the genes in each species to obtain super-sequences for phylogenetic tree building. Maximum-likelihood (ML) phylogenetic analysis with 1000 bootstrap repeats and discrete gamma distribution across sites was performed by PhyML3.0 software (Guindon et al., 2010). *C.elegans* was used as an outgroup. The CodeML model (Schabauer et al., 2012) in Phylogenetic Analysis by Maximum Likelihood (PAML) (Z. Yang, 2007) and a branch site model were applied to analyse the selective pressure of single-copy genes in each species. The functional annotation and enrichment analysis of the obtained rapidly evolving genes were carried out using GO and KEGG, respectively (Figure S5). On the basis of the phylogenetic tree, divergence time was estimated using MCMCTREE in the PAML package (Z. Yang, 2007). The TimeTree database (Hedges, Dudley, & Kumar, 2006) and divergence times were applied as the time controls, and the fossil calibration used in the evolutionary trees was derived. We assessed the convergence of the independent runs by a comparison of likelihood scores and model parameter estimates in TRACER v1.5 (Rambaut, Suchard, Xie, & Drummond, 2013).

The most recent common ancestor gene families of the 16 species were used in an analysis of expansion and contraction. OrthoMCL (version 2.0) was used to cluster the homologous groups of the 16 species. Comparisons of the expansion and contraction of orthologous gene families were performed by Computational Analysis of gene Family Evolution (CAFE v4.1) (De Bie, Cristianini, Demuth, & Hahn, 2006) using birth-death models to infer the process in the phylogeny of gene gain and loss.

2.7 | Identification of thechemosensory gene families

The chemosensory gene families were manually annotated using the NCBI BLAST program with *T. castaneum* sequences as queries (http://blast.ncbi.nlm.nih.gov/Blast.cgi). The protein sequences of olfactory receptors (ORs) and odorant binding proteins (OBPs) of*T. castaneum* (Tcas), *D . ponderosae* (Dpon) and *A. glabripennis* (Agla) were downloaded from GenBank (Andersson, Keeling, & Mitchell, 2019; Mitchell, Schneider, Schwartz, Andersson, & McKenna, 2020). The amino acid sequences were used to construct an ML tree. The phylogenetic tree was constructed in MEGA X10.0 using the ML method (Tamura et al., 2011) with a suitable evolutionary model. Finally, the tree was annotated using Evoview software (https://www.evolgenius.info/evolview) (Subramanian, Gao, Lercher, Hu, & Chen, 2019).

The tissue-specific expression profiles of *CforOBP4-36* were evaluated by real-time quantitative reverse transcription-polymerase chain reaction (qRT-PCR) analysis. cDNA was generated following the instructions

of the PrimeScript RT reagent Kit with gDNA Eraser (TaKaRa, Beijing, China). Three reference genes, namely, *ςφορβ-αςτιν*(GenBank accession MH 716465), *CforGAPDH* (GenBank accession MT 512411) and *CforUBE4A* (GenBank accession MT 512412), were used as the controls (Hua et al., 2021). The specific primers are listed in Table S1. qRT-PCR was performed using TB Green Premix Ex Taq II (TaKaRa, Beijing, China) in a CFX96 Real-Time PCR Detection System (Bio-Rad, Richmond, CA, USA). Thermal cycling was performed using the following parameters: initial denaturation at 94 °C for 30 s, followed by 40 cycles of 94 °C for 5 s and 60 °C for 30 s. According to the dilution concentration and the corresponding CT value, a standard curve was generated, and the amplification efficiency was calculated by the equation $E = 10^{-1/slope}$. The means and standard errors were calculated for three biological replicates with three technical replicates of each tissue and control. The relative Ct values were analysed using the 2^{-T} method.

2.8 | Expression and purification of recombinant CforOBP4-6 proteins

Recombinant CforOBP4-6 were expressed in *Escherichia coli* cells. Prokaryotic expression primers (Table S1) were designed for CforOBP4-6 (the signal peptide was removed) and contained the BamHI and EcoRI restriction sites. The positive clone was cultured in 5 mL of Luria-Bertani (LB) liquid medium containing kanamycin (50 μg/mL) and grown overnight at 37 °C with shaking (220 r/m). The culture was then diluted to 500 mL in LB medium and cultured at 37 °C. Recombinant protein expression was induced by the addition of isopropyl-β -D-1-thiogalactopyranoside (IPTG) at a final concentration of 1 mM when the OD (600 nm) of the culture reached 0.6 to 0.8. After 4 h of incubation at 28 °C with shaking (180 r/m), the cells were harvested by centrifugation (12000 r/m, 10 min, 4 °C) and sonicated in binding buffer (20 mM sodium phosphate, 0.5 M NaCl, 20 mM imidazole; pH 7.4). The recombinant CforOBP4-6 proteins were purified using a Ni-ion affinity chromatography column (HisTrap HP; GE Healthcare, Piscataway, NJ, USA). The His-tag was removed using enterokinase solution (2 U/mg) (Sigma St Louis, MO, USA) at 24 °C for 4 h. The purified protein was analysed by 15% sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) and then dialyzed using a HiTrap desalting column (GE Healthcare). The concentration of the protein was measured with the BCA Assay Kit (Sangong, Shanghai, China). The purified protein was stored at -80 °C.

2.9 | Fluorescence displacement binding assay

A fluorescence displacement binding assay was performed to determine the affinity of CforOBP4-6 for 102 volatiles (Table S2) according to the methods described in our previous study (Hua *et al.* , 2021). The ligand binding experiment was carried out on an F-4700 fluorescence spectrophotometer (Hitachi, Japan) using N-phenyl-1-naphthylamine (1-NPN) as a fluorescent probe with excitation at 337 nm, and emission spectra were recorded from 400 nm to 550 nm. 1-NPN and all the ligands were dissolved in methanol. The binding constants for 1-NPN were measured by adding 1-NPN to 2 μM CforOBP4-6 in 50 mM Tris-HCl buffer (pH 7.4) to achieve final concentrations ranging from 2 to 20 μM. The binding affinities of the ligands were measured using 1-NPN as a fluorescence probe with a stoichiometry of 1:1 (ligand:protein). The binding affinities of ligands were obtained from an average of three independent experiments with a stoichiometry of 1:1 (1-NPN and CforOBPs). The CforOBP4/1-NPN dissociation constant (Kd) and the curves were calculated from the relative Scatchard plots using GraphPad Prism 8 software (GraphPad, La Jolla, CA, USA). The binding affinity (Ki) of the competitors was calculated based on the IC50 values using the equation $Ki = [IC_{50}]/(1+[1\text{-NPN}]/K_{1\text{-NPN}})$, where [1-NPN] is the free concentration of 1-NPN and $K_{1\text{-NPN}}$ is the dissociation constant of the CforOBP4/1-NPN complex.

3 | Result

3.1 | Genome size evaluation

The genome size of*C. formicarius* was estimated by k-mer analyses of the Illumina DNA data. The average k-mer depth was 61 (i.e., the main peak) (Figure S1). The k-mer depth at twice that of the main peak was a repetitive peak, and the k-mer depth at half the depth of the main peak was heterozygous. The total number of k-mers obtained from the sequencing data was 24,314,168,078. After removing abnormal k-mers, a total of 22,494,390,501 k-mers were used to evaluate genome size and characteristics, and the *C. formicarius* genome

size was estimated to be 364.51 Mb. The k-mer distribution analysis showed that the repetitive sequence content was approximately 28.60% and that the heterozygosity was approximately 0.43% (Table S3). There was no obvious heterozygous peak. A quality check did not find plant and microbe contamination (Figure S2). These characteristics implied that the genome of *C. formicarius* is a simple genome, which is conducive to the construction of a genome map.

3.2 |Illumina and PacBio sequencing and genome assembly

To assemble and annotate a draft genome of *C. formicarius* , we used a hybrid sequencing approach to generate short and long reads from both the Illumina and PacBio platforms. Using Illumina PE genomic sequencing, approximately 27.64 Gb of raw data was obtained, and removal of low-quality reads and adapter sequences resulted in 27.40 Gb of clean data (Table S4). For long-read sequencing, after filtering raw reads, we obtained 2,884,459 reads and 34,139,672,427 bases (˜100x depth) on the PacBio platform. The statistical data showed that the mean subread length was 11.84 kb, the read N50 length was 18.75 kb, and the longest read was 89.71 kb. Finally, we obtained 27.40 Gb of short clean reads and 34.14 Gb of long clean reads (Table S5), which were combined to assemble the *C. formicarius*genome. A 338.84 Mb assembly was obtained for *C. formicarius* with a contig N50 length of 14.97 Mb and a longest contig length of 27.31 Mb (Table 1, Table S6). The*C. formicarius* assembly genome size is similar to the mean assembly size of the previously published coleopteran genomes (ranging from 110 to 850 Mb) (Hlerema *et al.* , 2017) and is comparable to the genome sizes of*Plastocerus angulosus* (367 Mbp) (Kusy, Motyka, Bocek, Vogler, & Bocak, 2018) and*Hycleus phaleratus* (308 Mbp) (Y. M. Wu, Li, & Chen, 2018).

To assess the assembly accuracy, the completeness of the draft genome was evaluated with single-copy orthologous genes using BUSCO, mapping the Illumina data to a reference genome, and analysis with CEGMA v2.5. The CEGMA v 2.5 analysis showed that 98.03% of the 458 conserved core genes of eukaryotes in the CEGMA database were completely detected, and 94.76% of the 248 highly conserved core genes of eukaryotes were found in the assembled genome (Table S8). Furthermore, 98.87% and 0.56% of the 1,054 highly conserved insect orthologues from BUSCO v3.0.1 were identified as complete and fragmented, respectively, in the assembly (Table S9). In total, approximately 98.01% reads and 96.04% proper reads were mapped to the assembled genome sequences (Table S7). These analyses indicated that the *C. formicarius* genome obtained was a high-quality assembly.

3.3 | Chromosome sequence assembly

We obtained 32.83 Gb of clean reads from the Hi-C fragment library after filtering, representing 97-fold coverage of the draft genome (338.84 Mb). In total, 80.04 Mb of unique mapped read pairs and 32.23 Mb of valid interaction pairs were generated (Table S10). After error correction with Illumina PE sequencing, PacBio long-read sequencing and Hi-C interaction maps, we obtained a final assembly that was 337.06 Mb in size, comprising 221 contigs and 154 scaffolds, with a contig N50 of 13.21 Mb and a scaffold N50 of 34.23 Mb. Finally, a total of 337.06 Mb of genome sequences, accounting for 99.42% of the assembled draft genome, was anchored to 11 pseudo-chromosomes. More importantly, 78 scaffolds, comprising 99.03% of the total sequence length, were ordered and oriented (Table S11, Figure 2). These results indicated that the assembled draft genome of*C. formicarius* had a high degree of continuity and completeness.

3.4 |Repeat annotation

In total, the *C . formicarius* genome was found to contain 157.51 Mb of repetitive sequences (approximately 46.49% of the assembled genome), of which 41.55% were transposable elements (TEs) (Table S12). DNA transposons accounted for 21.61% of the *C. formicarius* genome, and the most common classification assigned to these repetitive elements was terminal inverted repeat (TIR) (length of 64.78 Mb, 19.11% of assembly) (Table 2). The proportion of repetitive sequences in the *C. formicarius* genome was higher than that in other coleopteran genomes, such as the*Hypothenemus hampei* (2.7% of assembly) (Vega et al., 2015),*Nicrophorus vespilloides* (12.85% of assembly) (Cunningham et al., 2015),*Leptinotarsa decemlineata* (17% of assembly) (Schoville et al., 2018), *D .ponderosae* (17% and 23% of assembly for males and females, respectively) (Keeling *et al.* , 2013), *Hycleus cichorii* and*H. phaleratus* (22.73% and 13.47% of assembly, respectively) genomes, and

was similar to that in the*Pyrocoelia pectoralis* (44.88% of assembly) (Fu et al., 2017) and*Propylea japonica* (58.22%) (L. Zhang et al., 2020) genomes.

3.5 | Gene prediction and functional annotation We used three different methods to predict protein-coding genes in the*C. formicarius* genome, namely, ab initio, RNA-seq-based and homology-based methods. Then, we used EVM v1.1.1 software to integrate the prediction results. A total of 11,907 protein-coding genes were found in the *C. formicarius* genome (Table 3 and Table 4), which were divided into 9,291 gene families, 75 of which were unique gene families (Table S13). With the support of RNA-seq-based and homology-based gene prediction methods, the final prediction results showed 11,610 protein-coding genes, accounting for 97.51% of the total protein-coding genes (Figure S3), showing a good gene prediction effect on the *C. formicarius* genome. Functional annotation statistics showed that 11,469 genes (96.32% of the predicted genes) were assigned to corresponding putative functions (Table S14; Figure 3). Compared with other known coleopteran genomes, the number of genes in the *C. formicarius* genome was similar to that in *T. castaneum* (12,841),*O .taurus* (14,402), *D . ponderosae* (12,102), *A .glabripennis* (14,533), *O . borbonicus* (14,402), and*A . planipennis* (11,373) (Table S13). We also identified and annotated various non-coding RNAs sequences, including 102 rRNAs, 165 tRNAs and 40 miRNAs (Table S15).

3.6| Species phylogeny analysis The phylogenetic analysis showed that there were fewer species-specific genes in *C. formicarius* (223) than in the 15 other species of insects, except the coleopteran insect *O. borbonicus* (55) and hymenopteran insect *A. mellifera* (141) (Figure 4). The 11,907 one-to-one homologous genes from the gene family analysis were used to infer the phylogeny. Finally, 223 unigenes from *C. formicarius* were obtained, corresponding to 75 gene families. In total, 11,011 orthologues were identified, which could be clustered with the other 15 insects, including 896 unclusters (Table S13). Among them, the proportion of species-specific genes in the six coleopteran insect genomes ranged from 0.82% (*O. borbonicus*) to 7.38% (*O. taurus*). All bootstrap values of the nodes generated were above 90%, the majority being higher than 99%. *C. formicarius* and two other coleopteran insects clustered together. *C. formicarius* and *D. ponderosae* diverged from the common ancestor of *A. glabripennis*187.54 Mya, and the divergence time between *D. ponderosae* and*C. formicarius* was approximately 138.89 Mya (Figure 4). The coleopteran insects from the same order were clustered together and formed a clade, clearly sharing a common ancestor. The evolution time of the other insects was consistent with previous studies (McKenna *et al.* , 2016), and the differentiation time of *C. elegans* was the most primitive as an outgroup.

A total of 132,846 gene families in the most recent common ancestor of the 16 species were obtained by analysing the gene family expansion and contraction. The numbers of expanded and contracted gene families in*C. formicarius* were 31 and 28, respectively (Figure S5). Compared with *D . ponderosae* and *A . glabripennis* ,*C. formicarius* had 27 expanded and 16 contracted gene families, which demonstrated that the number of expanded genes in *C. formicarius* had increased significantly. This result indicated that*C. formicarius* may have experienced more duplication events than*D . ponderosae* . We found that these genes in *C. formicarius* were also the most abundant based on the multicopy homologous gene number (Figure S5). In addition, we performed GO and KEGG enrichment analyses of these expanded and contracted genes in the*C. formicarius* genome (Figure S6). We found lineage-specific expansion of genes related to the biosynthesis of chemosensory metabolites, which may affect the biosynthesis of olfaction-related proteins and improve the olfactory sensitivity of *C. formicarius*(Table S17).

3.7 | Identification of the chemosensory gene families

During interactions with the environment, *C. formicarius* gene families are probably involved in a variety of sensory processes, including in searching for food sources, locating mates and spawning sites, avoiding predators, exchange of information between individuals and socializing among groups (Hua *et al.* , 2021). To better study insect behaviour, the olfactory mechanism of *C. formicarius* was explored. We annotated a complete set of chemosensory genes in the existing *C. formicarius* genome (Table S16). As an obvious comparison for the coleopteran insect *C. formicarius* , the coleopteran insect*A. glabripennis* has 61 genes encoding OBPs, 132 genes encoding ORs, 234 genes encoding gustatory receptors (GRs), 72 genes encoding ionotropic receptors (IRs), 17 genes encoding chemosensory proteins (CSPs) and 4 genes encoding sensory

neuron membrane proteins (SNMPs) (McKenna *et al.* , 2016). We manually annotated 36 OBPs, 154 ORs, 46 GRs, 39 IRs, 13 CSPs, and 4 SNMPs in the *C. formicarius* genome (Table S16).

In general, the genome of *C. formicarius* encodes components similar to those of other weevils. The notable exceptions are the OR and OBP families, in which a total of 190 components (154 ORs and 36 OBPs) were found, indicating massive gene expansion in the *C. formicarius* genome. We compared the OR and OBP gene families involved in chemosensory activity between *C . formicarius* and*T.castaneum*, *D . ponderosae* and *A .glabripennis* (Figures 5 and 6). *C. formicarius* has 153 OR genes in addition to the highly conserved OR coreceptor CforOrco (Table S17: Figure 5). These include representatives of all seven subfamilies of beetle ORs except group 4/6 and follow the pattern of frequent paralogous radiation typical of insect chemoreceptors. One new lineage of ORs was identified and placed as group 8 in *C. formicarius*(CforOr91-150). Many *C. formicarius* ORs are in tandem arrays (Figure 5) and are derived from recent expansions. *C. formicarius*may thus harbour the larger identified insect OR repertoire, because there are 46 ORs in *A. planipennis* , 79 in *D. ponderosae* , 121 in *A. glabripennis* , and 270 in *T . castaneum*(Mitchell *et al.* , 2020). The large numbers of *C. formicarius* and *T . castaneum* ORs are thought to be due to current or past difficulties in findings hosts and food. As has been suggested for *Solenopsis invicta* (Wurm et al., 2011), the large number for *C. formicarius* may be due to the importance of chemical communication among weevils. OBPs constitute an essential family of genes that are also known to play roles in chemosensation in*Drosophila* (P. Xu, Atkinson, Jones, & Smith, 2005). The majority of identified OBPs comprise a large expansion of the minus-C subfamily, and the remaining genes encode the classic 6-cysteine motif and were placed alone or in a small radiation pattern. Five OBPs (CforOBP10, CforOBP11, CforOBP21, CforOBP22 and CforOBP24) were identified as members of the plus-C group and were identical to *T. castaneum*(TcasOBP6),*D. ponderosae* (DponOBP26) and *A. glabripennis* (AglaOBP15) (Figure 6).

3.8 | Tissue expression profile of *CforOBPs* in*C. formicarius*

To obtain initial insights into expression differences among tissue samples, we comparatively analysed the expression of *CforOBPs* in the main chemosensory tissue, the antennae, heads (the whole head capsule excluding the antennae), legs, thoraxes (excluding head and legs) and abdomens of males and females.

The expression of the *CforOBP* s was restricted to the main chemosensory tissues (antennae and heads) (Figure 7). The transcripts of 24 of the 33 *CforOBP* s were significantly enriched in antennae (*CforOBP4* , *CforOBP5* , *CforOBP7* , *CforOBP9* ,*CforOBP10* , *CforOBP11* , *CforOBP12* , *CforOBP14* ,*CforOBP15* , *CforOBP16* , *CforOBP17* , *CforOBP19* ,*CforOBP20* , *CforOBP21* , *CforOBP22* , *CforOBP23* ,*CforOBP24* , *CforOBP27* , *CforOBP28* , *CforOBP29* ,*CforOBP32* , *CforOBP34* , *CforOBP35* and*CforOBP36*), whereas three were enriched in the mouthparts (*CforOBP4* , *CforOBP5* and *CforOBP11*). Statistical analysis of the male and female antennal samples showed no significant difference; 13 of the 24 *CforOBP* s were significantly enriched in female antennae (*CforOBP5* , *CforOBP7* , *CforOBP10* ,*CforOBP14* , *CforOBP17* , *CforOBP19* , *CforOBP20* ,*CforOBP21* , *CforOBP22* , *CforOBP29* , *CforOBP32* ,*CforOBP34* and *CforOBP35*). Nevertheless, *CforOBP16* ,*CforOBP23* and *CforOBP28* showed more than fivefold over-expression in male antennae compared to female antennae (Figure 7). The fact that we found major and significant differences between males and females is consistent with anatomical data from the antennal lobe, where sexual dimorphism was found (Sutherland, 1986). Interestingly, seven of the *CforOBP* s were significantly enriched in the female abdomen (*CforOBP10* , *CforOBP13* , *CforOBP19* ,*CforOBP25* , *CforOBP30* , *CforOBP32* and*CforOBP33*). Most of the *CforOBP* s were expressed in the main chemosensory tissue, that is, antennae and heads, and only seven were significantly enriched in female antennae compared to female legs. Thus, these genes were most likely exclusively involved in chemosensory processing.

In our previous study, CforOBP1-3 were also enriched in antennae, whereas CforOBP1 was also detected in the abdomen and legs. Functional characteristic analysis results suggested that CforOBP1-3 are involved not only in the reception of sex pheromones but also in the behaviour of searching for host plants (Hua *et al.* , 2021). To further study the role of CforOBPs in olfactory recognition of *C .formicarius* , we performed functional analysis of CforOBP4-6 in vitro.

3.9 | Fluorescence binding assay

Recombinant *CforOBP4-6* proteins, expressed predominantly in antennae, were produced using a prokaryotic expression system. The prokaryotic expression vectors pET30a/CforOBP4-6 were successfully expressed in *E . coli* (Figure 8A). Purified recombinant mature proteins were obtained by cleaving the His-Tag using enterokinase. The purified recombinant proteins were examined by SDS-PAGE, as shown in Figure 8B.

To determine the binding specificity of antennae-enriched CforOBP4-6, 102 odorant compounds, including sex pheromones and host plant volatiles (Table S4), were chosen as ligands to characterize the binding properties of CforOBP4-6. The Kd values for CforOBP4-6 bound to 1-NPN were 3.295±0.151 μM, 3.072±0.1881 μM and 3.491±0.2524 μM, respectively (Figure 9A). Representative 1-NPN displacement curves were displayed in Figure 5A. Based on these curves, the median inhibitory concentration (IC50, displacement of more than 50% of 1-NPN) and the reciprocal values of the dissociation constant (Ki) were calculated (Figure 9).

The binding test results indicated that the Ki values of recombinant CforOBP4-6 with the sex pheromones were 1.564 ± 0.229 μM, 1.064 ± 0.221 μM and 1.351 ± 0.093 μM, respectively (Table 5), indicating strong binding affinities (Figure 9B-D). Among the 43 tested sweet potato volatiles, kaempferol-3-glucoside showed the highest binding affinity for CforOBP5, with a Ki of 0.289±0.026 μM, followed by the affinity of kaempferol 3,7,4'-trimethyl ether for CforOBP5, with a Ki of 1.370±0.028 μM, and those of tiliroside, 1-aminoanthracene and rhamnetin for CforOBP6, with Ki values of 1.690±0.114, 7.05±0.305 μM and 8.112±0.248 μM, respectively (Figure 5B-D).

4 | DISCUSSION

C . formicarius is the most important pest of sweet potato and invades many of the main areas of sweet potato production throughout the tropics and subtropics (Capinera, Jansson, & Raman, 1999; Hiroyoshi*et al.* , 2016; Reddy, Zhao, & Humber, 2014). With continued research on *C . formicarius* , its morphological, ecological and physiological characteristics have become clear. The mitochondrial genome (H. Yang & Li, 2019) and the transcriptome of *C . formicarius* (Bin*et al.* , 2017; Ma *et al.* , 2016) have been previously studied. However, abundant genomic resources and genome-wide molecular markers of *C . formicarius* and related species are still lacking. The genome sequence of *C . formicarius* provides a novel resource for Brentidae, many species of which are economically and ecologically important pests in agriculture. At the same time, it provides comparative data for the genome sequence and provides the basis for evolutionary research on coleopteran and other insects.

In this study, we report the first draft genome sequence of *C .formicarius* at the chromosome level using the Illumina and PacBio sequencing platforms and Hi-C technology. With recent developments in SMRT sequencing technology, chromosomal-level genome assembly using long-read sequencing strategies and Hi-C technology was also reported recently in insects (Y. Li, Park, Smith, & Moran, 2019; Liu et al., 2019; Meng et al., 2020; J. Yang et al., 2020). The high-quality genome that we assembled could provide an important resource for the molecular ecological development of *C .formicarius* and its related species. K-mer distribution analysis showed that the *C . formicarius*genome was small and homozygous compared with the genomes of other known coleopteran insects (Evans et al., 2018; Fu *et al.* , 2017; L. Zhang *et al.* , 2020). The low heterozygosity indicates the low level of the population genetic diversity of *C . formicarius* in the narrow feeding habitat, or it might be due to single-pair mating. Based on the k-mer depth distribution map, the estimated *C . formicarius* genome size was approximately 364.51Mb (Figures S1 and S2).

The *C . formicarius* genome size was determined to be approximately 338.84 Mb using SMRT assembly. Hi-C decodes the 3D structure of chromatin by detecting genome-wide DNA interactions (Belaghzal, Dekker, & Gibcus, 2017). Hi-C-assisted genome assembly is mainly based on two principles: the interaction between DNA fragments within chromosomes is greater than that between chromosomes, and the interaction signals and their linear distance are subject to a power law. (Dekker, Rippe, Dekker, & Kleckner, 2002; Flot, Marie-Nelly, & Koszul, 2015). Therefore, this technology can be used to determine the chromosomal locations of most of the sequences in the preliminary assembled draft genome and to identify the group, order and

orientation of these sequences and has been successfully used to assist genome assembly (Dudchenko et al., 2017; Rusk, 2014). Using Hi-C sequence data, 221 contigs were constructed on 11 chromosomes with a genome size of 337.06 Mb, with a scaffold N50 of 34.23 Mb and contig N50 of 14.97 Mb. In addition, 11,907 protein-coding genes were predicted in the genome, and 96.32% of the genes were functionally annotated through a search in public databases. To date, the *C . formicarius* assembly is the most comprehensive of all published coleopteran genomes, with the longest N50 of 14.97 Mb (L. Zhang *et al.* , 2020). The completeness and high quality of this assembly are comparable to those of other high-quality coleopteran genomes, providing a new paradigm for future assemblies of coleopteran genomes and even other insect genomes (Ando et al., 2018; Y. M. Wu *et al.* , 2018; L. Zhang *et al.* , 2020).

The *C . formicarius* genome assembly will facilitate evolutionary studies. The phylogenetic analysis of 16 insect species showed that the coleopteran insects diverged from the ancestor of dipteran and lepidopteran insects approximately 410 Mya, consistent with the ML phylogenetic analysis using 745 single-copy genes from 11 species representing five orders and one mite species (L. Zhang *et al.* , 2020). Coleopteran insects (six species) were clustered in the same lineages. The estimated divergence time between the ancestor of*T . castaneum* and *A. glabripennis* was approximately 223.47 Mya.*A. glabripennis* diverged from the ancestor of *D .ponderosae* and *C . formicarius* approximately 187.54 Mya;*C .formicarius* was sister to *D . ponderosae* , and*C . formicarius* diverged from *D . ponderosae*approximately 138.89 Mya, all of which are consistent with a previous phylogenetic analysis based on ML, with 523 orthologues from 15 insect species (McKenna *et al.* , 2016) and based on two ML methods (RAxML and IQ-TREE) and a Bayesian approach (ExaBayes), with 95 nuclear protein-coding genes from 373 beetle species (Mckenna et al., 2015; S. Q. Zhang et al., 2018).

Studies have shown that the male adults have a high affinity for sex pheromones released by the females (Coffelt *et al.* , 1978; Heath*et al.* , 1986), but the mechanisms of sex pheromone perception, especially ORs, have not been reported. Some key functional genes that appeared to expand in *C . formicarius* were also identified; these genes may play a role in environmental adaptation (N. Wu et al., 2019). These functional genes are associated with ORs, GRs and IRs, as well as OBPs, CSPs, and SNMPs, revealing their ability to interact with a diverse chemical environment (Keeling *et al.* , 2013; Richards*et al.* , 2008). Among these expanded genes, we identified ORs and OBPs that were inferred to mediate taste chemical neurotransmission and are potential targets for biological control (Bargmann, 2006; Sato et al., 2008; Tiwari, Karpe, & Sowdhamini, 2019; Wurm *et al.* , 2011). At the same time, the binding characteristics of CforOBP4-6 were identified. The fluorescent competitive binding assay results indicated that CforOBP4-6 had strong binding affinities for sex pheromones and other ligands.

These putative candidates will be particularly important for further research; for example, exploration of the function of ORs will help us to clarify the odorant recognition mechanism of insects and provide a theoretical basis screening more effective behavioural interference factors at the molecular level. The findings provide a new strategy for pest management by regulating odorant perception behaviour; therefore, ORs are believed to play an important role in the recognition of volatile compounds by insects (Fleischer, Pregitzer, Breer, & Krieger, 2018) and are frequently used to study the relationship among ecological specialization, adaptability and gene family evolution (Andersson*et al.* , 2019; Mitchell *et al.* , 2020). In the future, the gene families that exhibited expansion or contraction will be further studied. The functional verification of these candidate species will help in the study of the mechanism by which some invasive species adapt to other species and environments.

In summary, we constructed the first high-quality chromosome-level genome of *C . formicarius* , performed a comparative genomic analysis between this species and 15 other species, and found that OR and OBP gene families were expanded in *C .formicarius* . These datasets not only provide a wealth of information for studying the genetics and evolutionary mechanisms of this species but also provide very valuable resources for further study on the molecular mechanisms of stress resistance, allowing researchers to identify important functional genes and population genetic patterns of *C . formicarius* . A complete genome sequence will advance our understanding of the molecular mechanisms underlying the processes of tolerance to insecticides and abiotic and biotic stresses and will accelerate studies on population genetics, which will facilitate the

development of IPM of *C . formicarius* .

Acknowledgements

This research was supported by the China Agriculture Research System (CARS-10-B3 and CARS-10-C19), National Natural Science Foundation of China (31660627), Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), Science and Technology Development Foundation of Guangxi Academy of Agricultural Sciences (Guinongke2017JZ05 and 2018YM18), Natural Science Foundation of Guangxi Province (2016JJB130253), Guangxi Innovation Team Construction Project (nycytxgxcxtd-11-03), Postgraduate ResearchPractice Innovation Program of Jiangsu Province of China (KYCX18-2126) and Jiangsu Students' Platform for Innovation and Entrepreneurship Training Program (202010320111Y).

Data Accessibility

The raw genome sequencing reads and assembly were deposited in the NCBI Sequence Read Archive (SRA), with BioProject accession nos. PRJNA725324 and PRJNA725325 and BioSample accession no. SAMN18917023. The final chromosome assembly was submitted to NCBI Assembly under accession no. SRR14368409. Raw Illumina, PacBio, Hi-C and RNA-seq reads have been deposited into the NCBI SRA under accession nos. SRR14373957, SRR14368409, SRR14373634, and SRR14373956, respectively.

Author Contributions

Z.L., J.H., D.M. and T.C. designed the study; Y.H., X.G., J.S., H.L., Y.H. and Y.Li. collated the samples; J.H. and Y.F. performed the research; J.H., L.Z., X.D., Y.H., X.G. and J.S. analysed the data; J.H. wrote the manuscript; and Z.L. and L.Z. revised the manuscript. All authors approved the final manuscript.

Conflicts of Interest

The authors declare that they have no competing interests.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215* (3), 403-410. doi:10.1016/S0022-2836(05)80360-2

Andersson, M. N., Keeling, C. I., & Mitchell, R. F. (2019). Genomic content of chemosensory genes correlates with host range in wood-boring beetles (*Dendroctonus ponderosae* , *Agrilus planipennis* , and*Anoplophora glabripennis*). *BMC Genomics, 20* (690), 1-18. doi:10.21203/rs.2.11535/v1

Ando, T., Matsuda, T., Goto, K., Hara, K., Ito, A., Hirata, J., . . . Niimi, T. (2018). Repeated inversions within a *pannier* intron drive diversification of intraspecific colour patterns of ladybird beetles. *Nat Commun, 9* (3843), 1-13. doi:10.1038/s41467-018-06116-1

Andrews, S. (2014). FastQC A Quality Control tool for High Throughput Sequence Data.

Attwood, T. K., & Beck, M. E. (1994). PRINTS–a protein motif fingerprint database. *Protein Eng, 7* (7), 841-848. doi:10.1093/protein/7.7.841

Austin, D. F. (1988). The taxonomy, evolution and genetic diversity of sweet potatoes and related wild species. *Exploration Maintenance & Utilization of Sweet Potato Genetic Resources Rep Sweet Potato Planning Conf* .

Bairoch, A. (1991). PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res, 19 Suppl* , 2241-2245. doi:10.1093/nar/19.suppl.2241

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA, 6* (11), 1-6. doi:10.1186/s13100-015-0041-9

Bargmann, C. I. (2006). Comparative chemosensation from receptors to ecology. *Nature, 444* (7117), 295-301. doi:10.1038/nature05402

Belaghzal, H., Dekker, J., & Gibcus, J. H. (2017). Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. *Methods, 123* , 56-65. doi:10.1016/j.ymeth.2017.04.004

Bin, S. Y., Qu, M. Q., Pu, X. H., Wu, Z. Z., & Lin, J. T. (2017). Antennal transcriptome and expression analyses of olfactory genes in the sweetpotato weevil Cylas formicarius. *Sci Rep, 7* (1), 11073. doi:10.1038/s41598-017-11456-x

Birney, E., Clamp, M., & Durbin, R. (2004). GeneWise and Genomewise.*Genome Res, 14* (5), 988-995. doi:10.1101/gr.1865504

Blanco, E., Parra, G., & Guigo, R. (2007). Using geneid to identify genes. *Curr Protoc Bioinformatics, Chapter 4* (1), 1-28. doi:10.1002/0471250953.bi0403s18

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., . . . Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res, 31* (1), 365-370. doi:10.1093/nar/gkg095

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30* (15), 2114-2120. doi:10.1093/bioinformatics/btu170

Bouchard, P., Bousquet, Y., Davies, A. E., Alonso-Zarazaga, M. A., Lawrence, J. F., Lyal, C. H., . . . Smith, A. B. (2011). Family-group names in Coleoptera (Insecta). *Zookeys, 88* (1), 1-972. doi:10.3897/zookeys.88.807

Bovell-Benjamin, A. C. (2007). Sweet potato: a review of its past, present, and future role in human nutrition. *Adv Food Nutr Res, 52* (1), 1-59. doi:10.1016/S1043-4526(06)52001-7

Bru, C., Courcelle, E., Carrere, S., Beausse, Y., Dalmar, S., & Kahn, D. (2005). The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res, 33* (Database issue), D212-215. doi:10.1093/nar/gki034

Burge, C., & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J Mol Biol, 268* (1), 78-94. doi:10.1006/jmbi.1997.0951

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. *Nat Biotechnol, 31* (12), 1119-1125. doi:10.1038/nbt.2727

Capinera, J. L., Jansson, R. K., & Raman, K. V. (1999). Sweet potato pest management. a global perspective. *The Florida Entomologist, 82* (1), 125. doi:10.2307/3495845

Choi, J. H., Kijimoto, T., Snell-Rood, E., Tae, H., Yang, Y., Moczek, A. P., & Andrews, J. (2010). Gene discovery in the horned beetle Onthophagus taurus. *BMC Genomics, 11* , 703. doi:10.1186/1471-2164-11-703

Christiaens, O., Prentice, K., Pertry, I., Ghislain, M., Bailey, A., Niblett, C., . . . Smagghe, G. (2016). RNA interference: a promising biopesticide strategy against the African Sweetpotato Weevil *Cylas brunneus* . *Sci Rep, 6* (38836), 1-11. doi:10.1038/srep38836

Cockerham, K. L., Deen, O. T., Christian, M. B., & Newsom, L. T. (1954). The biology of the sweet potato weevil. *Technical Bulletin Louisiana Agricultural Experiment Station, 483* (10), 1-30.

Coffelt, J. A., Vick, K. W., Sower, L. L., & Mcclellan, W. T. (1978). Sex pheromone of the sweetpotato weevil, *Cylas formicarius*elegantulus laboratory bioassay and evidence for a multiple component system. *Environmental Entomology, 7* (5), 756-758. doi:10.1093/ee/7.5.756

Consortium, C. e. S. (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science, 282* (5396), 2012-2018. doi:10.1126/science.282.5396.2012

Cunningham, C. B., Ji, L., Wiberg, R. A., Shelton, J., McKinney, E. C., Parker, D. J., . . . Moore, A. J. (2015). The Genome and Methylome of a Beetle with Complex Social Behavior, Nicrophorus vespilloides (Coleoptera: Silphidae). *Genome Biol Evol, 7* (12), 3383-3396. doi:10.1093/gbe/evv194

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. doi:10.1093/bioinformatics/btl097

Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing chromosome conformation. *Science, 295* (5558), 1306-1311. doi:10.1126/science.1067799

Duan, J., Kuhn, K. L., Gibbs, R. A., Worley, K. C., Murali, S. C., Lee, S. L., . . . Richards, S. (2019). Agrilus planipennis genome assembly v1.0. . *Ag Data Commons* . doi:10.15482/USDA.ADC/1503806

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., . . . Aiden, E. L. (2017). De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds.*Science, 356* (6333), 92-95. doi:10.1126/science.aal3327

Edgar, R. C., & Myers, E. W. (2005). PILER: identification and classification of genomic repeats. *Bioinformatics, 21* (S1), i152-i158. doi:10.1093/bioinformatics/bti1003

Evans, J. D., McKenna, D., Scully, E., Cook, S. C., Dainat, B., Egekwu, N., . . . Huang, Q. (2018). Genome of the small hive beetle (*Aethina tumida* , Coleoptera: Nitidulidae), a worldwide parasite of social bee colonies, provides insights into detoxification and herbivory. *Gigascience, 7* (12), 1-16. doi:10.1093/gigascience/giy138

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., . . . Punta, M. (2014). Pfam: the protein families database. *Nucleic Acids Res, 42* (27), 1-9. doi:10.1093/nar/gkt1223

Fleischer, J., Pregitzer, P., Breer, H., & Krieger, J. (2018). Access to the odor world: olfactory receptors and their role for signal transduction in insects. *Cell Mol Life Sci, 75* (3), 485-508. doi:10.1007/s00018-017-2627-5

Flot, J. F., Marie-Nelly, H., & Koszul, R. (2015). Contact genomics: scaffolding and phasing (meta)genomes using chromosome 3D physical signatures. *FEBS Lett, 589* (20 Pt A), 2966-2974. doi:10.1016/j.febslet.2015.04.034

Fu, X., Li, J., Tian, Y., Quan, W., Zhang, S., Liu, Q., . . . Hu, J. (2017). Long-read sequence assembly of the firefly *Pyrocoelia pectoralis* genome. *Gigascience, 6* (12), 1-7. doi:10.1093/gigascience/gix112

Gelbart, W. M. (1992). The return of the fly. *Science, 257* (5075), 1421-1422. doi:10.1126/science.257.5075.1421

Gough, J., & Chothia, C. (2002). SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res, 30* (1), 268-272. doi:10.1093/nar/30.1.268

Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A., & Enright, A. J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res, 34* (1), D140-D144. doi:10.1093/nar/gkj112

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0.*Syst Biol, 59* (3), 307-321. doi:10.1093/sysbio/syq010

Gunter, N. L., Oberprieler, R. G., & Cameron, S. L. (2016). Molecular phylogenetics of Australian weevils (Coleoptera: Curculionoidea): exploring relationships in a hyperdiverse lineage through comparison of independent analyses. *Austral Entomology, 55* (2), 217-233. doi:10.1111/aen.12173

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K., Jr., Hannick, L. I., . . . White, O. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies.*Nucleic Acids Res, 31* (19), 5654-5666. doi:10.1093/nar/gkg770

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., . . . Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol, 9* (1), 1-22. doi:10.1186/gb-2008-9-1-r7

Haft, D. H., Selengut, J. D., & White, O. (2003). The TIGRFAMs database of protein families. *Nucleic Acids Res, 31* (1), 371-373. doi:10.1093/nar/gkg128

Han, Y., & Wessler, S. R. (2010). MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res, 38* (22), 1-8. doi:10.1093/nar/gkq862

Hardee, D. D., Jones, G. D., & Adams, L. C. (1999). Emergence, movement, and host plants of boll weevils (Coleoptera : Curculionidae) in the Delta of Mississippi. *Journal of Economic Entomology, 92* (1), 130-139. doi:DOI 10.1093/jee/92.1.130

Heath, R. R., Coffelt, J. A., Sonnet, P. E., Proshold, F. I., Dueben, B., & Tumlinson, J. H. (1986). Identification of sex pheromone produced by female sweetpotato weevil,*Cylas formicarius elegantulus*(Summers). *J Chem Ecol, 12* (6), 1489-1503. doi:10.1007/BF01012367

Hedges, S. B., Dudley, J., & Kumar, S. (2006). TimeTree: a public knowledge-base of divergence times among organisms.*Bioinformatics, 22* (23), 2971-2972. doi:10.1093/bioinformatics/btl505

Hiroyoshi, S., Kohama, T., & Reddy, G. V. P. (2016). Age-related sperm production, transfer, and storage in the sweet potato weevil,*Cylas formicarius* (Fabricius) (Coleoptera: Curculionidae).*Journal of Insect Behavior, 29* (6), 689-707. doi:10.1007/s10905-016-9590-0

Hlerema, I., Laurie, S., & Eiasu, B. (2017). Preliminary observations on use of *Beauveria bassiana* for the control of the sweet potato weevil (*Cylas* sp.) in South Africa. *Open Agriculture, 2* (1), 595-599. doi:10.1515/opag-2017-0063

Hua, J., Pan, C., Huang, Y., Li, Y., Li, H., Wu, C., . . . Li, Z. (2021). Functional characteristic analysis of three Odorant-binding proteins from the sweet potato weevil (*Cylas formicarius*) in the perception of sex pheromones and host plant volatiles. *Pest Manag Sci, 77* (1), 300-312. doi:10.1002/ps.6019

Industry, D. O. P. (2016). Primitive weevils of Florida (Insecta: Coleoptera: Brentidae: Brentinae). *J Entomology Nematology, 87* (4), 1-4.

International Aphid Genomics, C. (2010). Genome sequence of the pea aphid Acyrthosiphon pisum. *PLoS Biol, 8* (2), e1000313. doi:10.1371/journal.pbio.1000313

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., . . . Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics, 30* (9), 1236-1240. doi:10.1093/bioinformatics/btu031

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res, 110* (1-4), 462-467. doi:10.1159/000084979

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E. P., Rivas, E., Eddy, S. R., . . . Petrov, A. I. (2018). Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res, 46* (D1), D335-D342. doi:10.1093/nar/gkx1038

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res, 28* (1), 27-30. doi:10.1093/nar/28.1.27

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res, 30* (14), 3059-3066. doi:10.1093/nar/gkf436

Keeling, C. I., Yuen, M. M., Liao, N. Y., Docking, T. R., Chan, S. K., Taylor, G. A., . . . Bohlmann, J. (2013). Draft genome of the mountain pine beetle, *Dendroctonus ponderosae* Hopkins, a major forest pest. *Genome Biol, 14* (3), 1-19. doi:10.1186/gb-2013-14-3-r27

Keilwagen, J., Wenk, M., Erickson, J. L., Schattat, M. H., Grau, J., & Hartung, F. (2016). Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res, 44* (9), 1-12. doi:10.1093/nar/gkw092

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

Korada, R. R., & Mukherjee, A. (2012). Management of sweet potato weevil, *Cylas formicarius* : a world review. *Fruit, Vegetable and Cereal Science and Biotechnology, 6* (S1), 79-92.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res, 27* (5), 722-736. doi:10.1101/gr.215087.116

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics, 5* (59), 1-9. doi:10.1186/1471-2105-5-59

Kusy, D., Motyka, M., Bocek, M., Vogler, A. P., & Bocak, L. (2018). Genome sequences identify three families of Coleoptera as morphologically derived click beetles (Elateridae). *Scientific Reports, 8* (1). doi:ARTN 17084

10.1038/s41598-018-35328-0

Kyereko, W. T., Hongbo, Z., Amoanimaa-Dede, H., Meiwei, G., & Yeboah, A. (2019). The major sweet potato weevils; management and control: a review. *Entomology, Ornithology & Herpetology: Current Research, 8* (2), 1-9. doi:10.35248/2171-0983.8.218

Lees, J., Yeats, C., Perkins, J., Sillitoe, I., Rentzsch, R., Dessailly, B. H., & Orengo, C. (2012). Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis. *Nucleic Acids Res, 40* (s1), D465-D471. doi:10.1093/nar/gkr1181

Letunic, I., Copley, R. R., Schmidt, S., Ciccarelli, F. D., Doerks, T., Schultz, J., . . . Bork, P. (2004). SMART 4.0: towards genomic data integration. *Nucleic Acids Res, 32* (s1), D142-D144. doi:10.1093/nar/gkh088

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25* (14), 1754-1760. doi:10.1093/bioinformatics/btp324

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

Li, Y., Park, H., Smith, T. E., & Moran, N. A. (2019). Gene family evolution in the pea aphid based on chromosome-level genome assembly.*Mol Biol Evol, 36* (10), 2143-2156. doi:10.1093/molbev/msz138

Lima, T., Auchincloss, A. H., Coudert, E., Keller, G., Michoud, K., Rivoire, C., . . . Bairoch, A. (2009). HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. *Nucleic Acids Res, 37* (s1), D471-D478. doi:10.1093/nar/gkn661

Liu, Q., Guo, Y., Zhang, Y., Hu, W., Li, Y., Zhu, D., . . . Zhou, X. N. (2019). A chromosomal-level genome assembly for the insect vector for Chagas disease, *Triatoma rubrofasciata* . *Gigascience, 8* (8), 1-8. doi:10.1093/gigascience/giz089

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res, 25* (5), 955-964. doi:10.1093/nar/25.5.955

Ma, J., Wang, R., Li, X., Gao, B., & Chen, S. (2016). Transcriptome and gene expression analysis of *Cylas formicarius* (Coleoptera: Brentidae) during different development stages. *J Insect Sci, 16* (1), 1-11. doi:10.1093/jisesa/iew053

Majoros, W. H., Pertea, M., & Salzberg, S. L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders.*Bioinformatics, 20* (16), 2878-2879. doi:10.1093/bioinformatics/bth315

Marcais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.*Bioinformatics, 27* (6), 764-770. doi:10.1093/bioinformatics/btr011

Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., . . . Bryant, S. H. (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res, 39* (1), D225-D229. doi:10.1093/nar/gkq1189

McConnell, J. S., & Hossner, L. R. (1991). pH-dependent adsorption isotherms of glyphosate [Erratum to document cited in CA103(25):208751y]. *Journal of Agricultural and Food Chemistry, 39* (4), 824-824. doi:10.1021/jf00004a043

McKenna, D. D., Scully, E. D., Pauchet, Y., Hoover, K., Kirsch, R., Geib, S. M., . . . Richards, S. (2016). Genome of the Asian longhorned beetle (*Anoplophora glabripennis*), a globally significant invasive species, reveals key functional and evolutionary innovations at the beetle-plant interface. *Genome Biol, 17* (227), 1-18. doi:10.1186/s13059-016-1088-8

Mckenna, D. D., Wild, A. L., Kanda, K., Bellamy, C. L., Beutel, R. G., Caterino, M. S., . . . Farrell, B. D. (2015). The beetle tree of life reveals that Coleoptera survived end-Permian mass extinction to diversify during the Cretaceous terrestrial revolution. *Systematic Entomology, 40* (4), 835-880. doi:10.1111/syen.12132

Meng, F., Liu, Z., Han, H., Finkelbergs, D., Jiang, Y., Zhu, M., . . . Cai, J. (2020). Chromosome-level genome assembly of *Aldrichina grahami* , a forensically important blowfly. *Gigascience, 9* (3), 1-12. doi:10.1093/gigascience/giaa020

Meyer, J. M., Markov, G. V., Baskaran, P., Herrmann, M., Sommer, R. J., & Rödelsperger, C. (2016). Draft Genome of the Scarab BeetleOryctes borbonicuson La Réunion Island. *Genome Biology and Evolution, 8* (7), 2093-2105. doi:10.1093/gbe/evw133

Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., . . . Thomas, P. D. (2005). The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res, 33* (Database issue), D284-288. doi:10.1093/nar/gki078

Mitchell, R. F., Schneider, T. M., Schwartz, A. M., Andersson, M. N., & McKenna, D. D. (2020). The diversity and evolution of odorant receptors in beetles (Coleoptera). *Insect Mol Biol, 29* (1), 77-91. doi:10.1111/imb.12611

Nagano, T., Lubling, Y., Yaffe, E., Wingett, S. W., Dean, W., Tanay, A., & Fraser, P. (2015). Single-cell Hi-C for genome-wide detection of chromatin interactions that occur simultaneously in a single cell.*Nature Protocols, 10* (12), 1986-2003. doi:10.1038/nprot.2015.127

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

Ondiaka, S., Maniania, N. K., Nyamasyo, G. H. N., & Nderitu, J. H. (2008). Virulence of the entomopathogenic fungi *Beauveria bassiana* and *Metarhizium anisopliae* to sweet potato weevil*Cylas puncticollis* and effects on fecundity and egg viability.*Annals of Applied Biology, 153* (1), 41-48. doi:10.1111/j.1744-7348.2008.00236.x

Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes.*Bioinformatics, 23* (9), 1061-1067. doi:10.1093/bioinformatics/btm071

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., & Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc, 11* (9), 1650-1667. doi:10.1038/nprot.2016.095

Pittendrigh, B. R., Clark, J. M., Johnston, J. S., Lee, S. H., Romero-Severson, J., & Dasch, G. A. (2006). Sequencing of a new target genome: the Pediculus humanus humanus (Phthiraptera: Pediculidae) genome project. *J Med Entomol, 43* (6), 1103-1111. doi:10.1603/0022-2585(2006)43[1103:soantg]2.0.co;2

Price, A. L., Jones, N. C., & Pevzner, P. A. (2005). De novo identification of repeat families in large genomes.*Bioinformatics, 21* (S1), i351-i358. doi:10.1093/bioinformatics/bti1018

Rambaut, A., Suchard, M. A., Xie, D., & Drummond, A. J. (2013). Tracer v1.5. *Available online at: http://beast.bio.ed.ac.uk/Tracer* .

Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., . . . Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.*Cell, 159* (7), 1665-1680. doi:10.1016/j.cell.2014.11.021

Reddy, G. V. P., Zhao, Z. H., & Humber, R. A. (2014). Laboratory and field efficacy of entomopathogenic fungi for the management of the sweetpotato weevil, Cylas formicarius (Coleoptera: Brentidae).*Journal of Invertebrate Pathology, 122* , 10-15. doi:10.1016/j.jip.2014.07.009

Richards, S., Gibbs, R. A., Weinstock, G. M., Brown, S. J., Denell, R., Beeman, R. W., . . . Bucher, G. (2008). The genome of the model beetle and pest *Tribolium castaneum* . *Nature, 452* (7190), 949-955. doi:10.1038/nature06784

Rosenfeld, J. A., Reeves, D., Brugler, M. R., Narechania, A., Simon, S., Durrett, R., . . . Mason, C. E. (2016). Genome assembly and geospatial phylogenomics of the bed bug Cimex lectularius. *Nat Commun, 7* , 10164. doi:10.1038/ncomms10164

Ruan, J., & Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nat Methods, 17* (2), 155-158. doi:10.1038/s41592-019-0669-3

Rusk, N. (2014). Genomes in 3D improve one-dimensional assemblies.*Nat Methods, 11* (1), 5. doi:10.1038/nmeth.2795

Sato, K., Pellegrino, M., Nakagawa, T., Nakagawa, T., Vosshall, L. B., & Touhara, K. (2008). Insect olfactory receptors are heteromeric ligand-gated ion channels. *Nature, 452* (7190), 1002-1006. doi:10.1038/nature06850

Schabauer, H., Valle, M., Pacher, C., Stockinger, H., Stamatakis, A., Robinson-Rechavi, M., . . . Salamin, N. (2012). *SlimCodeML: An Optimized Version of CodeML for the Branch-Site Model.* Paper presented at the Parallel & Distributed Processing Symposium Workshops & Phd Forum. IEEE Computer Society.

Schon, K., & Skuhrovec, J. (2016). A new species of the genus*Corimalia* Gozis, 1885 (Coleoptera: Brentidae: Nanophyinae) from the Caucasus. *Zootaxa, 4169* (3), 571-578. doi:10.11646/zootaxa.4169.3.9

Schoville, S. D., Chen, Y. H., Andersson, M. N., Benoit, J. B., Bhandari, A., Bowsher, J. H., . . . Richards, S. (2018). A model species for agricultural pest genomics: the genome of the Colorado potato beetle, Leptinotarsa decemlineata (Coleoptera: Chrysomelidae). *Sci Rep, 8* (1), 1931. doi:10.1038/s41598-018-20154-1

Sequencing Consortium, T. H. G. (2006). Erratum: Insights into social insects from the genome of the honeybee Apis mellifera. *Nature, 444* (7118), 512-512. doi:10.1038/nature05400

Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C. J., Vert, J. P., . . . Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol, 16* (1), 259-270. doi:10.1186/s13059-015-0831-x

She, R., Chu, J. S., Wang, K., Pei, J., & Chen, N. (2009). GenBlastA: enabling BLAST to identify homologous gene sequences. *Genome Res, 19* (1), 143-149. doi:10.1101/gr.082081.108

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. doi:10.1093/bioinformatics/btv351

Staden, R. (1980). A new computer method for the storage and manipulation of DNA gel reading data. *Nucleic Acids Res, 8* (16), 3673-3694. doi:10.1093/nar/8.16.3673

Stanke, M., Steinkamp, R., Waack, S., & Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res, 32* (Web Server issue), W309-312. doi:10.1093/nar/gkh379

Subramanian, B., Gao, S., Lercher, M. J., Hu, S., & Chen, W. H. (2019). Evolview v3: a webserver for visualization, annotation, and management of phylogenetic trees. *Nucleic Acids Res, 47* (W1), W270-W275. doi:10.1093/nar/gkz357

Sutherland, J. A. (1986). A review of the biology and control of the sweetpotato weevil *Cylas formicarius* (Fab). *Tropical Pest Manag, 32* (4), 304-315. doi:10.1080/09670878609371084

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.*Mol Biol Evol, 28* (10), 2731-2739. doi:10.1093/molbev/msr121

Tang, S., Lomsadze, A., & Borodovsky, M. (2015). Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res, 43* (12), 1-10. doi:10.1093/nar/gkv227

Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics, Chapter 4* (1), 1-14. doi:10.1002/0471250953.bi0410s25

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., . . . Koonin, E. V. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research, 29* (1), 22-28. doi:DOI 10.1093/nar/29.1.22

Terrapon, N., Li, C., Robertson, H. M., Ji, L., Meng, X., Booth, W., . . . Liebig, J. (2014). Molecular traces of alternative social organization in a termite genome. *Nat Commun, 5* , 3636. doi:10.1038/ncomms4636

Tiwari, V., Karpe, S. D., & Sowdhamini, R. (2019). Topology prediction of insect olfactory receptors. *Curr Opin Struct Biol, 55* (1), 194-203. doi:10.1016/j.sbi.2019.05.014

Treangen, T. J., & Salzberg, S. L. (2012). Erratum: Repetitive DNA and next-generation sequencing: computational challenges and solutions.*Nature Reviews Genetics, 13* (2), 146-146. doi:10.1038/nrg3164

Vega, F. E., Brown, S. M., Chen, H., Shen, E., Nair, M. B., Ceja-Navarro, J. A., . . . Pain, A. (2015). Draft genome of the most devastating insect pest of coffee worldwide: the coffee berry borer, Hypothenemus hampei. *Sci Rep, 5* , 12525. doi:10.1038/srep12525

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., . . . Earl, A. M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One, 9* (11), 1-14. doi:10.1371/journal.pone.0112963

Wang, X., Fang, X., Yang, P., Jiang, X., Jiang, F., Zhao, D., . . . Kang, L. (2014). The locust genome provides insight into swarm formation and long-distance flight. *Nat Commun, 5* , 2957. doi:10.1038/ncomms3957

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., . . . Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nat Rev Genet, 8* (12), 973-982. doi:10.1038/nrg2165

Wolfe, G. W. (1991). The origin and dispersal of the pest species of*Cylas* with a key to the pest species groups of the world. *J Sweet Potato Pest Management A Global Perspective* , 13-43.

Wu, C. H., Nikolskaya, A., Huang, H. Z., Yeh, L. S. L., Natale, D. A., Vinayaka, C. R., . . . Barker, W. C. (2004). PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Res, 32* (s1), D112-D114. doi:10.1093/nar/gkh097

Wu, N., Zhang, S., Li, X., Cao, Y., Liu, X., Wang, Q., . . . Zhan, S. (2019). Fall webworm genomes yield insights into rapid adaptation of invasive species. *Nat Ecol Evol, 3* (1), 105-115. doi:10.1038/s41559-018-0746-5

Wu, Y. M., Li, J., & Chen, X. S. (2018). Draft genomes of two blister beetles *Hycleus cichorii* and *Hycleus phaleratus* .*Gigascience, 7* (3), 1-7. doi:10.1093/gigascience/giy006

Wurm, Y., Wang, J., Riba-Grognuz, O., Corona, M., Nygaard, S., Hunt, B. G., . . . Keller, L. (2011). The genome of the fire ant *Solenopsis invicta* . *Proc Natl Acad Sci U S A, 108* (14), 5679-5684. doi:10.1073/pnas.1009690108

Xia, Q., Zhou, Z., Lu, C., Cheng, D., Dai, F., Li, B., . . . Biology Analysis, G. (2004). A draft sequence for the genome of the domesticated silkworm (Bombyx mori). *Science, 306* (5703), 1937-1940. doi:10.1126/science.1102210

Xu, P., Atkinson, R., Jones, D. N., & Smith, D. P. (2005). Drosophila OBP LUSH is required for activity of pheromone-sensitive neurons.*Neuron, 45* (2), 193-200. doi:10.1016/j.neuron.2004.12.031

Xu, Z., & Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res, 35* (1), W265-W268. doi:10.1093/nar/gkm286

Yang, H., & Li, Y. (2019). Complete mitochondrial genome of *Cylas formicarius* (Coleoptera: Brentidae) from China. *Mitochondrial DNA Part B-Resources, 4* (1), 1241-1242. doi:10.1080/23802359.2019.1591247

Yang, J., Moeinzadeh, M. H., Kuhl, H., Helmuth, J., Xiao, P., Haas, S., . . . Vingron, M. (2017). Haplotype-resolved sweet potato genome traces back its hexaploidization history. *Nat Plants, 3* (9), 696-703. doi:10.1038/s41477-017-0002-z

Yang, J., Wan, W., Xie, M., Mao, J., Dong, Z., Lu, S., . . . Li, X. (2020). Chromosome-level reference genome assembly and gene editing of the dead-leaf butterfly *Kallima inachus* . *Mol Ecol Resour, 20* (4), 1080-1092. doi:10.1111/1755-0998.13185

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood.*Mol Biol Evol, 24* (8), 1586-1591. doi:10.1093/molbev/msm088

Zhang, L., Li, S., Luo, J., Du, P., Wu, L., Li, Y., . . . Cui, J. (2020). Chromosome-level genome assembly of the predator *Propylea japonica* to understand its tolerance to insecticides and high temperatures. *Mol Ecol Resour, 20* (1), 292-307. doi:10.1111/1755-0998.13100

Zhang, S. Q., Che, L. H., Li, Y., Dan, L., Pang, H., Slipinski, A., & Zhang, P. (2018). Evolutionary history of Coleoptera revealed by extensive sampling of genes and species. *Nat Commun, 9* (205), 1-12. doi:10.1038/s41467-017-02644-4

Figure and Table Legends

Figure 1. *Cylas formicarius*. a-g indicates different phenotypes. (a) Egg. (b) Larva. (c) Pupa. (d) Female. (e) Male. (f) The antennal sensilla showed strong sexual dimorphism in females and males. (g) Death feigning.

Figure 2 Genome-wide chromosomal contact matrix for *Cylas formicarius* showing interactions among the 11 chromosomes. The number of log2 links was calculated as the interaction frequency distribution of Hi-C links between and within chromosomes. The colour key of the heatmap ranging from light yellow to dark red indicates the frequency of Hi-C interaction links from low to high.

Figure 3 Venn diagram of functional annotation based on five databases. NR, KOG, GO, KEGG and TrEMBL.

Figure 4 Phylogenetic trees and gene orthology of *Cylas formicarius* and six other coleopteran insects and model insects. Node values show the divergence times from the present (million years ago, Mya). 1:1:1 (single-copy orthologous genes in common gene families); N:N:N (multiple-copy orthologous genes in common gene families); Specific (genes from a unique gene family from each species); Other (genes that do not belong to any abovementioned orthologous categories); Uncluster (genes that do not cluster with any families). Insect-specific genes are those that are present in only the 16 insect species. *Tribolium castaneum* (Richards *et al.* , 2008),*Onthophagus taurus* (Choi *et al.* , 2010),*Dendroctonus ponderosae* (Keeling *et al.* , 2013),*Anoplophora glabripennis* (McKenna *et al.* , 2016),*Oryctes borbonicus* (Meyer *et al.* , 2016),*Agrilus planipennis* (Duan *et al.* , 2019),*Bombyx mori* (Xia *et al.* , 2004), *Apis mellifera* (Sequencing Consortium, 2006), *Locustamigratoria* (Wang *et al.* , 2014), *Drosophilamelanogaster* (Gelbart, 1992), *Acyrthosiphon pisum*(International Aphid Genomics, 2010), *Pediculus humanus*(Pittendrigh *et al.* , 2006), *Cimex lectularius*(Rosenfeld *et al.* , 2016), *Zootermopsis nevadensis*(Terrapon *et al.* , 2014), *Caenorhabditis elegans*(Consortium, 1998).

Figure 5. Phylogenetic tree of the odorant receptor (OR) family. The receptor sequences included were from *Cylas formicarius* (Cfor, red), *Dendroctonus ponderosae* (Dpon, blue),*Anoplophora glabripennis* (Agla, green), *Tribolium castaneum* (Tcas, violet) and *Onthophagus taurus* (Otau, brown). The coloured arcs indicate the seven major coleopteran OR groups. To reduce tree size, the massively expanded coleopteran-specific OR lineages in former OR groups 1-7 are represented here by 5 ORs each. The sources of sequence data and explanation of receptor suffixes are detailed in Table S11.

Figure 6. Phylogenetic tree of CforOBPs and OBPs from other insect species. The OBPs originated from *Tribolium castaneum*(Tcas), *Anoplophora chinensis* (Achi), *Sitophilus zeamais*(Szea), and *Dendroctonus ponderosae* (Dpon).

Figure 7. Expression profile of *CforOBPs*. Heatmap showing the relative expression level of CforOBP4-36 as log2 in different tissues (adult antennae, heads (missing antennae but including mouth parts), legs, thoraxes and abdomens). The expression levels are represented by log2 red with high expression levels. The asterisks mark statistically significant differentially expressed genes compared to the expression in female legs. The black asterisks represent upregulation, and the red asterisks represent downregulation (p-values are * < 0.05; ** < 0.01; *** < 0.001).

Figure 8. A, Double restriction enzyme digestion of the pET/CforOBP4-6 prokaryotic expression vector; B, SDS-PAGE analysis of CforOBP4–6 purification. M, protein molecular marker; A 1,3,5, the pET/CforOBP4-6 vector was not digested; A 2,4,6, the pET/CforOBP4-6 vector was digested. B 1,4,7, total protein expression in *E. coli*pET/CforOBP4-6 was not induced by IPTG; B 2,5,8, total protein expression in *E. coli* pET/CforOBP4-6 was induced by IPTG; 3,6,9, purified pET/CforOBP4-6.

Figure 9. Competitive binding curves of CforOBP4-6 with various odorant compounds. (A) Affinity of CforOBP4-6 for the fluorescent probe N-phenyl-1-naphthylamine (1-NPN). Binding of CforOBP4 (B), CforOBP5 (C), and CforOBP6 (D) to ligands. (E) Comparison of the binding ability [calculated as 1/Ki (reciprocals of the dissociation constants) values] of these three proteins with (Z)-3-dodecen-1-yl(E)-2-butenoate and 22 ligands that exhibited significant affinity.

Table 1 Summary of assembly results for *Cylas formicarius*obtained by different methods

Table 2 Statistics of repeat elements

Table 3 Statistics of gene prediction results based on three methods

Table 4 Summary of the consensus gene set of the *Cylas formicarius* genome

Table 5. Binding affinities of the CforOBP4-6 proteins to 102 chemical compounds

Supplemental information

Figure S1. Distribution frequency of 19-mers in the *Cylas formicarius* genome.

Figure S2. Distribution of *Cylas formicarius* reads in the top 13 species according to the comparisons of 10000 NGS reads and the NT library.

Figure S3 Distribution of predicted genes by three methods: ab initio, homology-based method, and RNA-seq.

Figure S4 Gene functional annotation of *Cylas formicarius* was performed by alignment to the Gene Ontology (GO) database (a), eukaryotic orthologous groups of proteins (KOG) database (c), nucleotide collection (nr/nt) (b) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (d).

Figure S5 Phylogenetic tree of 16 insect species and the sizes of expanded and contracted gene families. *C. formicarius* ,*D . ponderosae* , *O . borbonicus* , *T .castaneum* , *A . planipennis* , *A .glabripennis* , *O . taurus* , *A .mellifera* , *C . lectularius* , *D .melanogaster* , *B . mori* , *P . humanus* ,*L . migratoria manilensis* , *C . elegans* ,*Z . nevadensis* and *A . pisum* .

Figure S6 Functional annotation and enrichment analysis of the obtained rapidly evolving genes were carried out using GO (a) and KEGG (b). (a) GO enrichment analysis of expanded gene families of *C. formicarius* . Bars are subdivided to represent different GO terms. (b) KEGG pathway enrichment analysis was performed for gene family expansion of *C. formicarius* . The graph depicts the most highly enriched pathways.

Table S1. Primer pairs used for cloning, prokaryotic expression and gene expression analysis using qRT-PCR.

Table S2. Purity and source of standard chemical compounds.

Table S3. Statistical information of the 19-kmer analysis of the *Cylas formicarius* genome.

Table S4. Summary of sequence reads generated using the Illumina system.

Table S5 Summarized sequence reads derived from the PacBio system.

Table S6 Length distribution of PacBio clean reads.

Table S7 Summary of Illumina reads aligned to the *Cylas formicarius* genome assembly.

Table S8 CEGMA evaluation of the *Cylas formicarius* genome assembly based on 458 core eukaryotic genes (CEGs) and 248 highly conserved CEGs.

Table S9 BUSCO evaluation for the *Cylas formicarius* genome assembly based on 1054 core eukaryotic genes.

Table S10. Summary of Hi-C data for error correction and chromosome assembly.

Table S11. Hi-C libraries for chromosome-level assembly.

Table S12. Statistics of repeat sequences.

Table S13 Statistics of the comparison of gene sets of *Cylas formicarius* and 15 other insect species.

Table S14 Statistics of functional annotation of predicted genes.

Table S15 Statistics of the predicted non-coding RNA.

Table S16. Comparison of the chemoreceptor and detoxification superfamilies of various species.

Table S17. Sequences used for building phylogenetic trees.

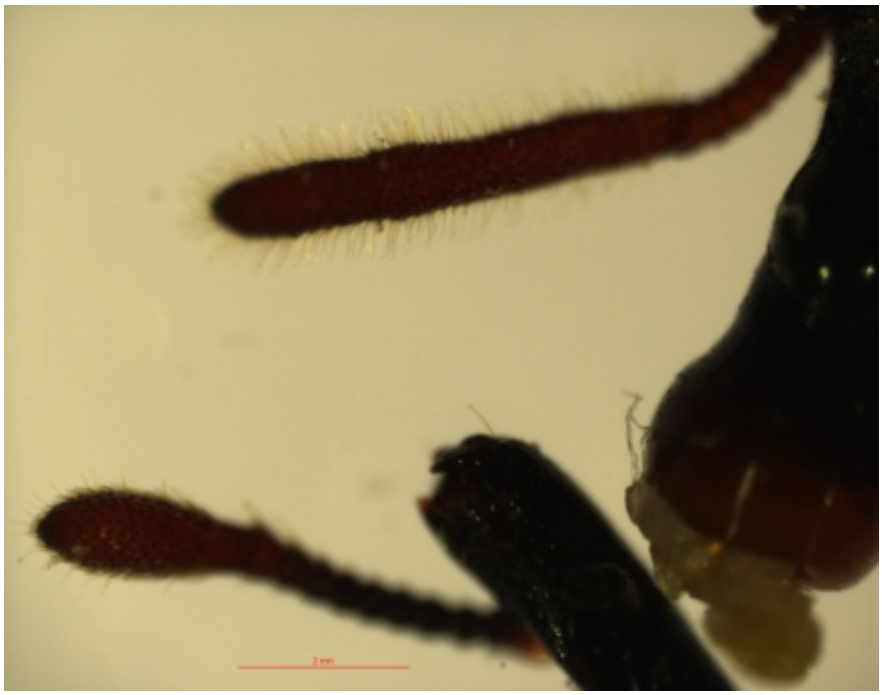


Figure 1. *Cylas formicarius*. a-g indicates different phenotypes. (a) Egg. (b) Larva. (c) Pupa. (d) Female. (e) Male. (f) The antennal sensilla showed strong sexual dimorphism in females and males. (g) Death feigning.

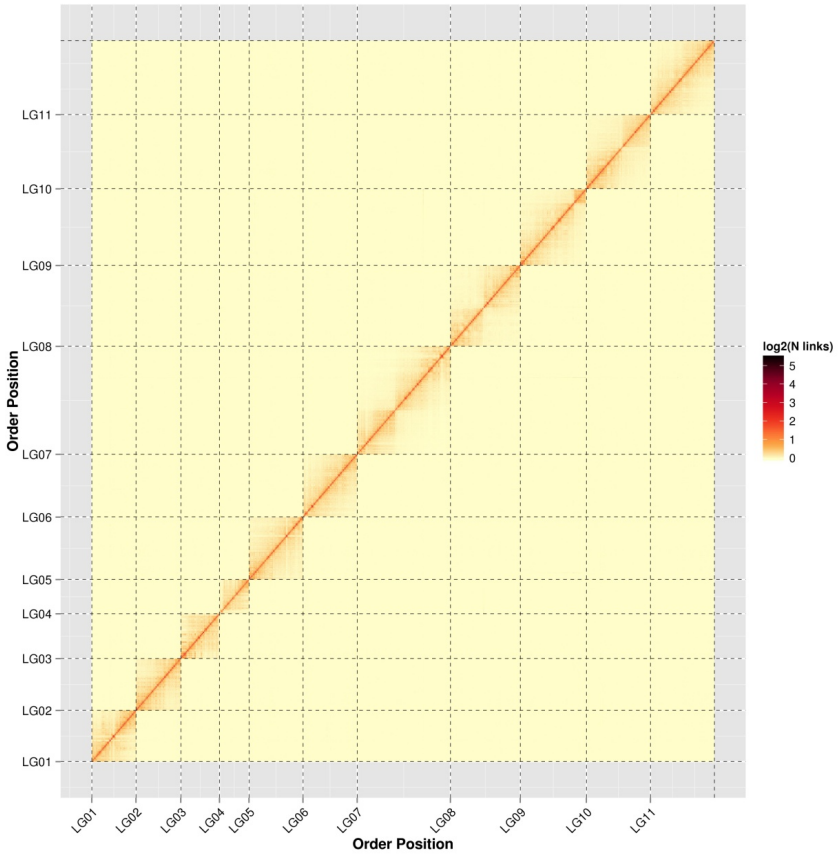


Figure 2 Genome-wide chromosomal contact matrix for *Cylas formicarius* showing interactions among the 11 chromosomes. The number of log2 links was calculated as the interaction frequency distribution of Hi-C links between and within chromosomes. The colour key of the heatmap ranging from light yellow to dark red indicates the frequency of Hi-C interaction links from low to high.

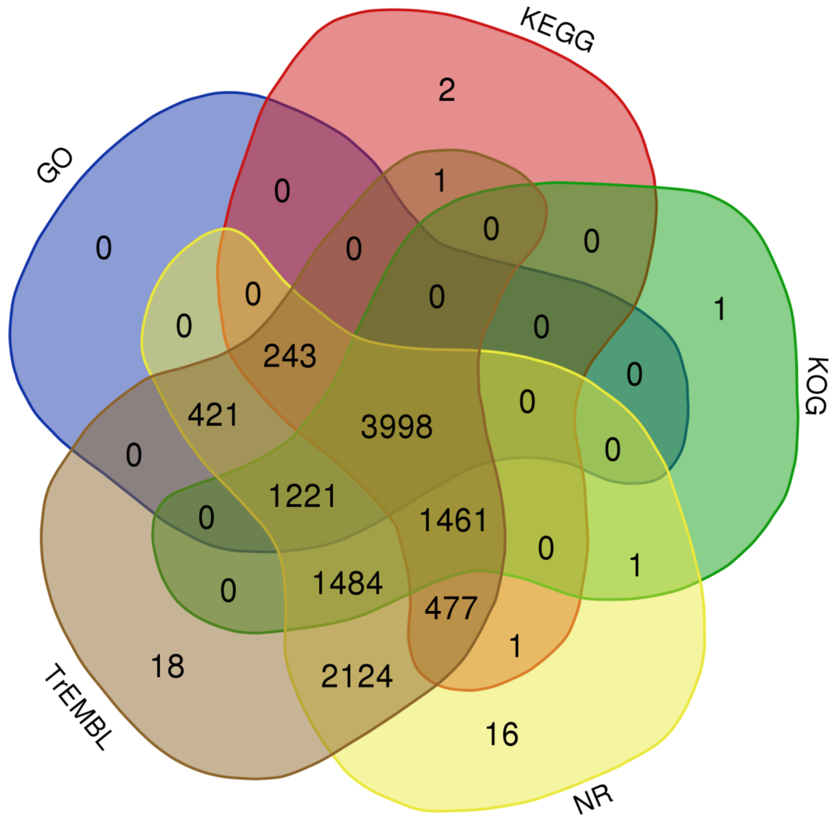


Figure 3 Venn diagram of functional annotation based on five databases. NR, KOG, GO, KEGG and TrEMBL.

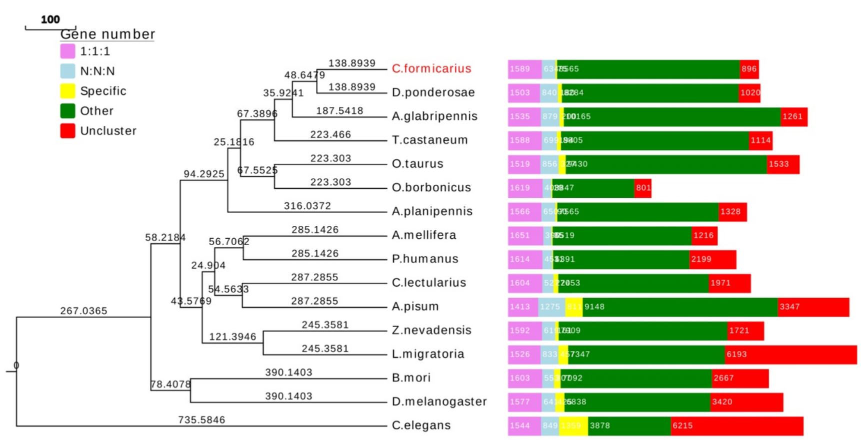


Figure 4 Phylogenetic trees and gene orthology of *Cylas formicarius* and six other coleopteran insects and model insects. Node values show the divergence times from the present (million years ago, Mya). 1:1:1 (single-copy orthologous genes in common gene families); N:N:N (multiple-copy orthologous

genes in common gene families); Specific (genes from a unique gene family from each species); Other (genes that do not belong to any abovementioned orthologous categories); Uncluster (genes that do not cluster with any families). Insect-specific genes are those that are present in only the 16 insect species. *Tribolium castaneum* (Richards *et al.* , 2008),*Onthophagus taurus* (Choi *et al.* , 2010),*Dendroctonus ponderosae* (Keeling *et al.* , 2013),*Anoplophora glabripennis* (McKenna *et al.* , 2016),*Oryctes borbonicus* (Meyer *et al.* , 2016),*Agrilus planipennis* (Duan *et al.* , 2019),*Bombyx mori* (Xia *et al.* , 2004), *Apis mellifera* (Sequencing Consortium, 2006), *Locustamigratoria* (Wang *et al.* , 2014), *Drosophilamelanogaster* (Gelbart, 1992), *Acyrthosiphon pisum*(International Aphid Genomics, 2010), *Pediculus humanus*(Pittendrigh *et al.* , 2006), *Cimex lectularius*(Rosenfeld *et al.* , 2016), *Zootermopsis nevadensis*(Terrapon *et al.* , 2014), *Caenorhabditis elegans*(Consortium, 1998).

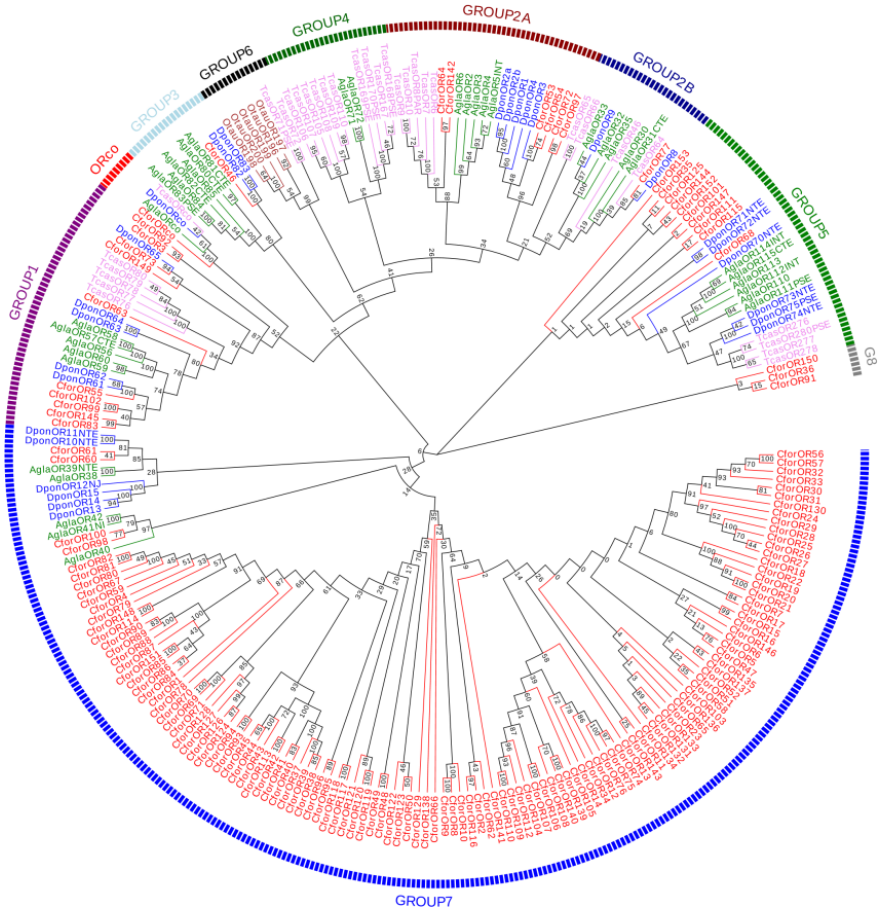


Figure 5. Phylogenetic tree of the odorant receptor (OR) family. The receptor sequences included were from *Cylas formicarius* (Cfor, red), *Dendroctonus ponderosae* (Dpon, blue),*Anoplophora glabripennis* (Agla, green), *Tribolium castaneum* (Tcas, violet) and *Onthophagus taurus* (Otau, brown). The coloured arcs indicate the seven major coleopteran OR groups. To reduce tree size, the massively expanded coleopteran-specific OR lineages in former OR groups 1-7 are represented here by 5 ORs each. The sources of sequence data and explanation of receptor suffixes are detailed in Table S11.

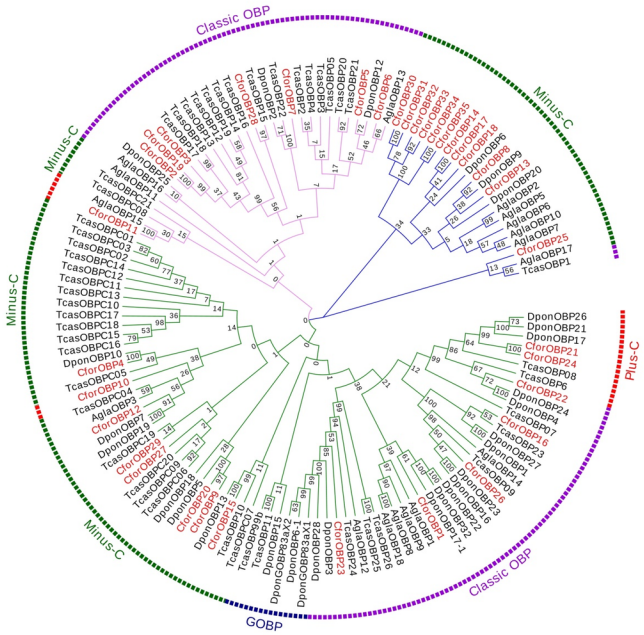


Figure 6. Phylogenetic tree of CforOBPs and OBPs from other insect species. The OBPs originated from *Tribolium castaneum*(Tcas), *Anoplophora chinensis* (Achi), *Sitophilus zeamais*(Szea), and *Dendroctonus ponderosae* (Dpon).

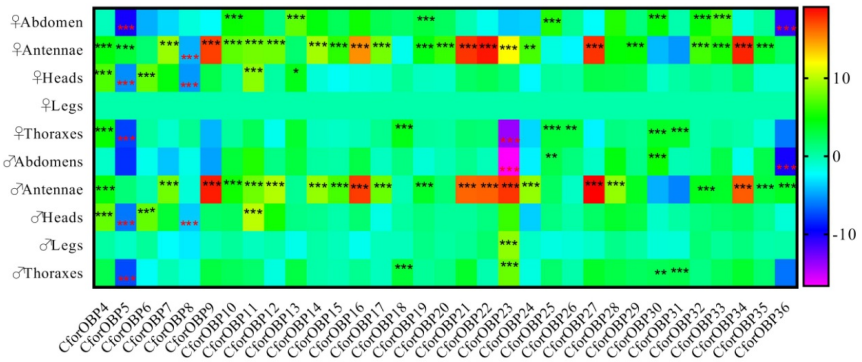


Figure 7. Expression profile of *CforOBPs*. Heatmap showing the relative expression level of CforOBP4-36 as log2 in different tissues (adult antennae, heads (missing antennae but including mouth parts), legs, thoraxes and abdomens). The expression levels are represented by log2 red with high expression levels. The asterisks mark statistically significant differentially expressed genes compared to the expression in female legs. The black asterisks represent upregulation, and the red asterisks represent downregulation (p-values are * < 0.05; ** < 0.01; *** < 0.001).

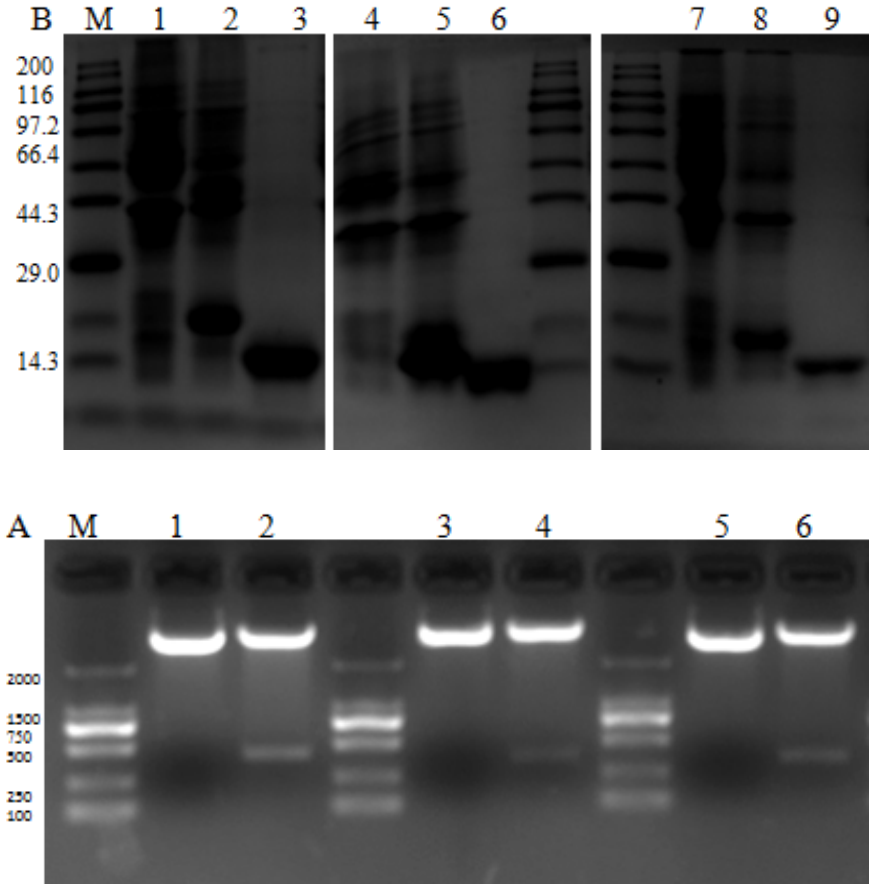
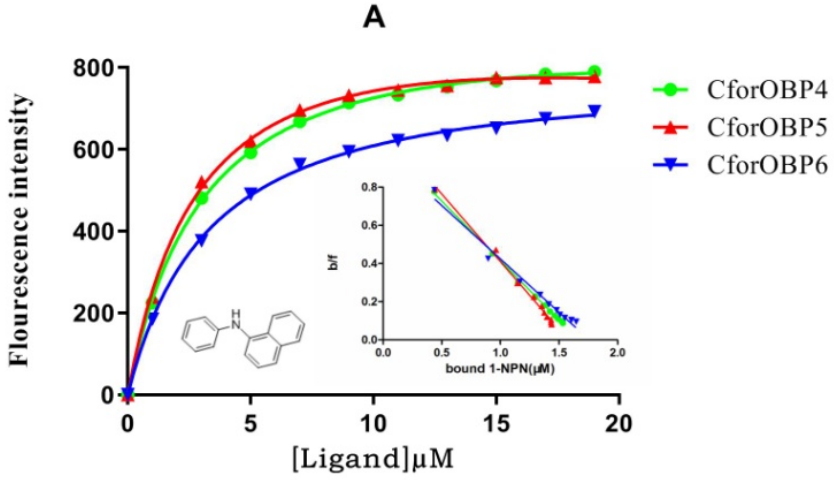
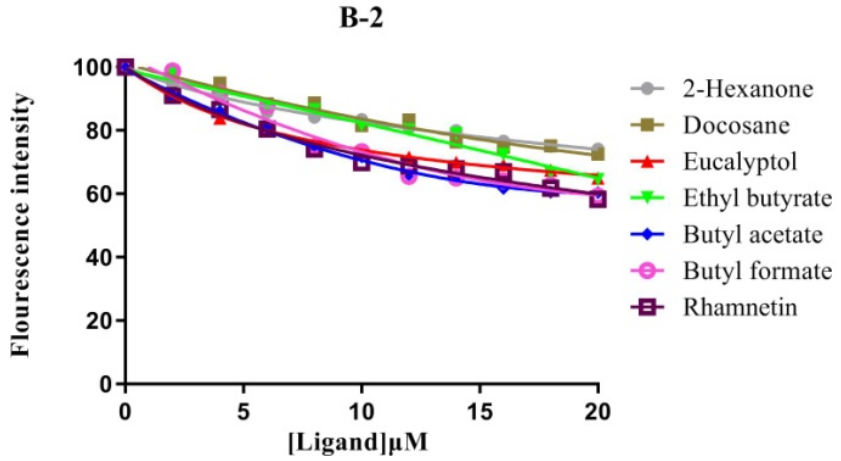
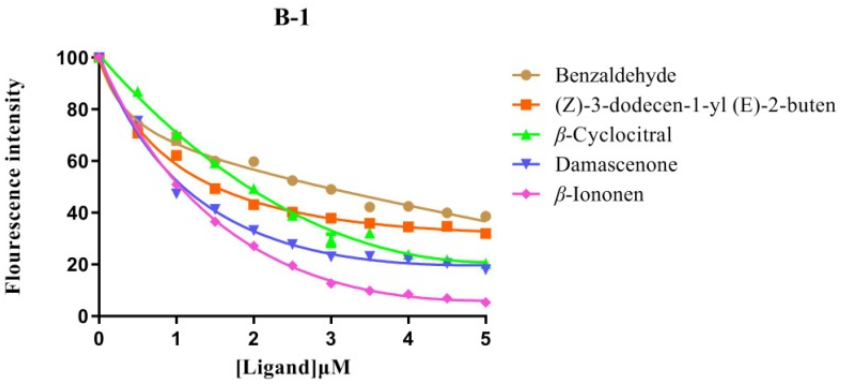
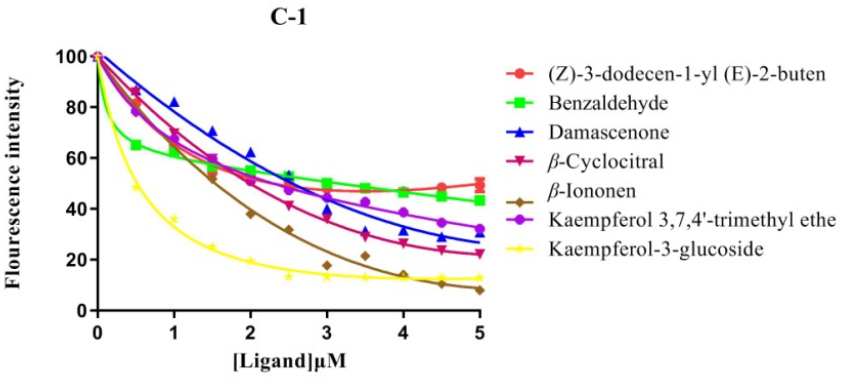


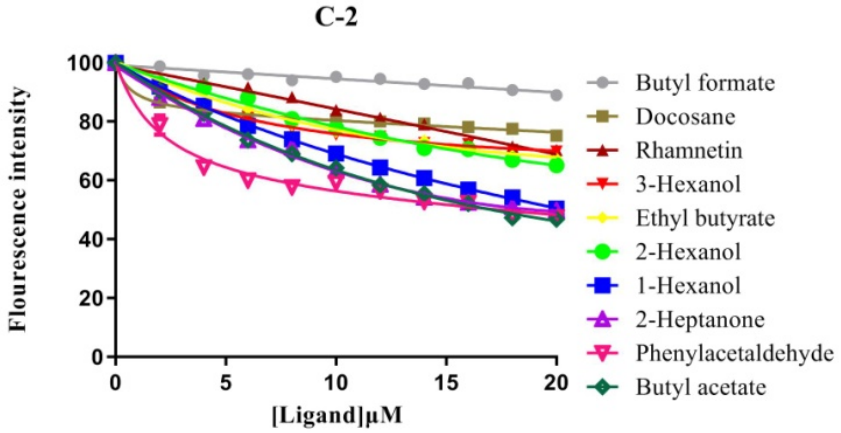
Figure 8. A, Double restriction enzyme digestion of the pET/CforOBP4-6 prokaryotic expression vector; B, SDS-PAGE analysis of CforOBP4–6 purification. M, protein molecular marker; A 1,3,5, the pET/CforOBP4-6 vector was not digested; A 2,4,6, the pET/CforOBP4-6 vector was digested. B 1,4,7, total protein expression in *E. coli*pET/CforOBP4-6 was not induced by IPTG; B 2,5,8, total protein expression in *E. coli* pET/CforOBP4-6 was induced by IPTG; 3,6,9, purified pET/CforOBP4-6.

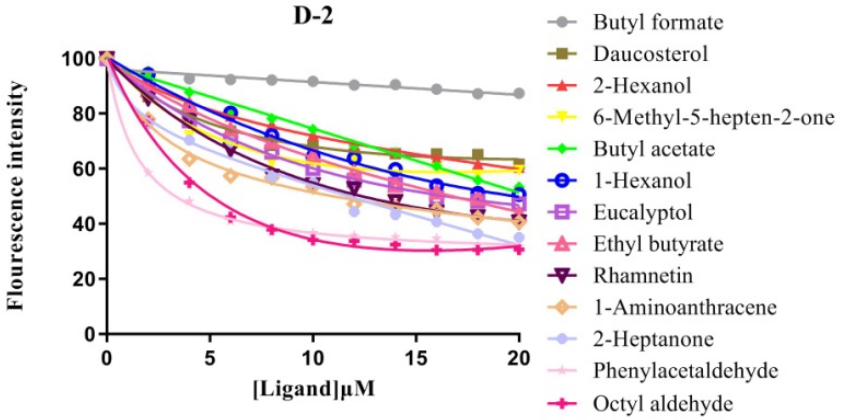


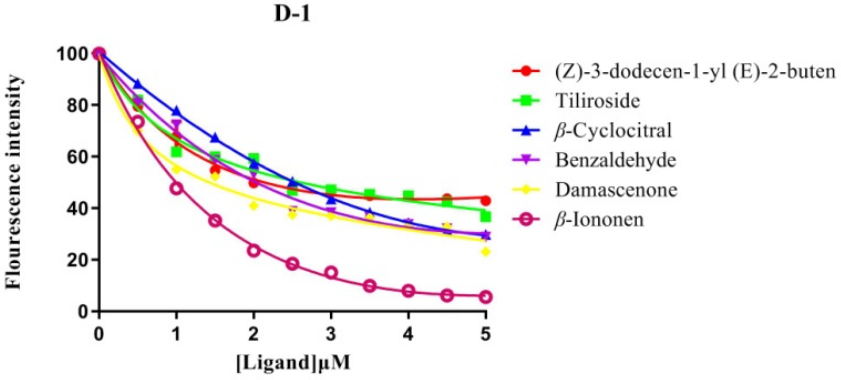












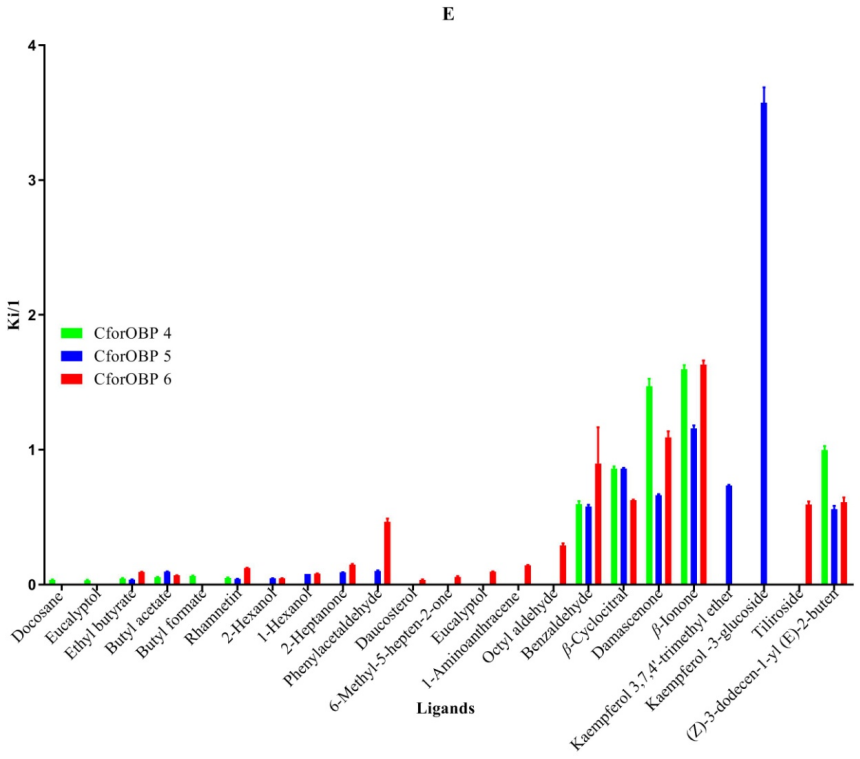


Figure 9. Competitive binding curves of CforOBP4-6 with various odorant compounds. (A) Affinity of CforOBP4-6 for the fluorescent probe N-phenyl-1-naphthylamine (1-NPN). Binding of CforOBP4 (B), CforOBP5 (C), and CforOBP6 (D) to ligands. (E) Comparison of the binding ability [calculated as 1/Ki (reciprocals of the dissociation constants) values] of these three proteins with (Z)-3-dodecen-1-yl(E)-2-butenoate and 22 ligands that exhibited significant affinity.

Table 1 Summary of assembly results of different methods for *Cylas formicarius*

Strategies	Contig	Scaffold
SMRT assembly		
Genome size (bp)	34,139,672,427	-
Total number	2,884,459	-
Max length (bp)	89,710	-
N50 size (bp)	18,725	-
Contig N90 (bp)	4,519,195	-
Hi-C scaffolding and gap filling		
Genome size (bp)	339,045,436	339,038,736
Total number	154	221
Max length (bp)	49,944,160	27,328,994
N50 size (bp)	34,231,294	13,216,042
Contig N90 (bp)	20,733,671	2,902,866

Table 2 Statistics of repeat elements

Classes	Type	Number	Length	Rate(%)
RNA		491713	103073651	30.4
	DIRS	6732	2301757	0.68
	LARD	55504	10642243	3.14
	LINE	129821	27870928	8.22
	LTR/Copia	2014	801356	0.24
	LTR/Gypsy	9889	4594013	1.35
	LTR/Unknown	5977	3263649	0.96
	PLE	278835	61058986	18.01
	SINE	444	86177	0.03
	TRIM	1693	2070686	0.61
	Unknown	804	321973	0.09
DNA		367692	73273338	21.61
	Crypton	3940	1124157	0.33
	Helitron	2751	558071	0.16
	MITE	495	73898	0.02
	Maverick	29512	7703503	2.27
	TIR	323939	64780365	19.11
	Unknown	7055	1086541	0.32
	PotentialHostGene	476	153271	0.05
	SSR	658	362661	0.11
	Unknown	45612	8740806	2.58
	Total	906151	157507930	46.46

Table 3 Statistics of gene prediction results based on three methods

Method	Software and gene set	Gene number
Ab initio	Genscan	14,203
	Augustus	10,593
	GlimmerHMM	36,128
	GeneID	8,104
	SNAP	22,127
Homology-based	*Anoplophora glabripennis*	12,478
	Dendroctonus ponderosae	10,499
	Oryctes borbonicus	8,839
	Tribolium castaneum	10,489
	Drosophila melanogaster	7,434
RNA-seq	PASA	17,697
	TransDecoder	23,437
	GeneMarkS-T	35,199
Integration	EVM	11,907

Table 4 Summary of the consensus gene set of the *Cylas formicarius* genome

Software		EVM
Gene number	Gene number	11,907
Gene length (bp)	Gene length (bp)	119,607,661
Average gene length (bp)	Average gene length (bp)	10,045

Software		EVM
Exon length (bp)	Exon length (bp)	26,779,793
Average exon length (bp)	Average exon length (bp)	2249
Intron length (bp)	Intron length (bp)	92,827,868
Average intron length (bp)	Average intron length (bp)	7,796

Table 5. Binding affinities of the CforOBP4-6 proteins to 102 chemical compounds.

Potential ligands [a]
(Z)-3-dodecen-1-yl (E)-2-buten [b]
Benzaldehyde
β-Cyclocitral
Damascenone
β-Ionone
Kaempferol 3,7,4'-trimethyl ether [c]
Kaempferol -3-glucoside[c]
Tiliroside [c]
2-Hexanone
Docosane [c]
Eucalyptol
Ethyl butyrate
Butyl acetate
Butyl formate
Rhamnetin [c]
3-Hexanol
2-Hexanol
1-Hexanol
2-Heptanone
Phenylacetaldehyde
Daucosterol [c]
6-Methyl-5-hepten-2-one
Eucalyptol
1-Aminoanthracene [c]
Octyl aldehyde
a: More potential ligands were tested, but the remaining 77 potential ligands did not bind any of the CforOBP4-6 proteins.

Hosted file

`Supplemental Information-2021.6.5.doc` available at https://authorea.com/users/421871/articles/527720-chromosome-level-genome-assembly-of-the-sweet-potato-weevil-cylas-formicarius-fabricius-coleoptera-brentidae-and-functional-characteristics-of-cforobp4-6