Chromosome-level genome of the peach fruit moth *Carposina sasakii* (Lepidoptera: Carposinidae) provides a resource for evolutionary studies on moths

Li-Jun Cao 1, Wei Song 1, Lei Yue 1, Shao-Kun Guo 1, Jin-Cui Chen 1, Ya-Jun Gong 1, Ary Anthony Hoffmann 2, Shu-Jun Wei 1, *" 

1. Institute of Plant and Environmental Protection, Beijing Academy of Agriculture and Forestry Sciences, 9 Shuguanghuayuan Middle Road, Haidian District, Beijing 100097, China 
2. School of BioSciences, Bio21 Institute, University of Melbourne, Parkville, Victoria, Australia 

* Corresponding author:
Short title: Chromosome-level genome of Carposina sasakii
Abstract

The peach fruit moth (PFM), *Carposina sasakii* Matsumura, is a major phytophagous orchard pest widely distributed across Northeast Asia. Here, we report the chromosome-level genome for the PFM, representing the first genome for the family Carposinidae, from the lepidopteran superfamily Copromorphoidea. The genome was assembled into 404.83 Mb sequences using PacBio long-read and Illumina short-read sequences, including 275 contigs, with a contig N50 length of 2.62 Mb. All contigs were assembled into 31 linkage groups assisted by the Hi-C technique, including 30 autosomes and a Z chromosome. BUSCO analysis showed that 98.3% of genes were complete and 0.4% of genes were fragmented, while 1.3% of genes were missing in the assembled genome. In total, 21,697 protein-coding genes were predicted, of which 84.80% were functionally annotated. Because of the importance of diapause triggered by photoperiod in PFM, five circadian genes in the PFM as well as in the other related species were annotated, and potential genes related to diapause and photoperiodic reaction were also identified from transcriptome sequencing. In addition, manual annotation of detoxification gene families was undertaken and showed a higher number of glutathione S-transferase (GST) gene in PFM than in most other lepidopterans, in contrast to a lower number of uridine diphosphate (UDP)-glycosyltransferase (UGT) gene, carboxyl/cholinesterases (CCE) gene and cytochrome P450 monooxygenase (P450) gene, suggesting different detoxication pathways in this moth. The high-quality genome provides a resource for comparative evolutionary studies of this moth and its relatives within the context of radiations across Lepidoptera.

Keywords: Insect; *de novo* assembly; annotation; detoxification; diapause

Introduction

The peach fruit moth (PFM), *Carposina sasakii* Matsumura (Lepidoptera: Carposinidae, superfamily Copromorphoidea), is a major phytophagous orchard pest of fruit such as apple, pear, peach, apricot and jujube from the families of Rosaceae and Rhamnaceae (Fig. 1). The hatched larvae directly bore into fruit to feed, causing losses in fruit production. PFM is one
of the most severe borers on deciduous fruit in northeast Asia. It is also considered a potential risk to fruit production in most parts of the world, although PFM is currently restricted to northeast Asia and far east Russia (Kwon et al., 2018; Wang et al., 2017).

One possible reason for the currently restricted distribution of PFM is its sensitivity to environmental factors. PFM has evolved diapause to cope with cold winter conditions and to synchronize its phenology with host plants (Toshima et al., 1961). Both long-day and short-day photoperiods induce diapause in the last instar of PFM larvae, resulting in a diapausing cocoon (Hua et al., 1998; Huang et al., 1976). The life cycle of PFM can be univoltine or bivoltine, depending on photoperiods encountered and environmental factors like humidity (Chiba & Kobayashi, 1985; Kim et al., 2000; Sato & Ishitani, 1976). Temperature also affects the occurrence of PFM through effects on the development rate and the emergence of the overwintering generation from diapause (Kim et al., 2001; Zhang et al., 2016).

The effects of environmental factors as well as photoperiod on the life history of PFM provides an opportunity to investigate the genomic basis of adaptation to temperate environments in the Copromorphoidea superfamily and across Lepidoptera more generally. Candidate genes involved in climatic adaptation could then also be investigated at the geographic level, given that a combination of mtDNA and microsatellite variation indicates strong genetic differentiation among geographical populations of the PFM across its native range in China (Wang et al., 2017). Although adaptation to climate changes has been found in lepidopteran species (Fält-Nardman et al., 2016; Quezada García et al., 2015; van Asch et al., 2013; Yamanaka et al., 2008), genes underlying climate adaptation remain to be identified (Hoffmann, 2017). Analysis of genomic variation among populations may help to identify putative loci and processes under selection (Hoffmann & Sgro, 2011) and predict the evolutionary adaptation under climate changes (Hoffmann & Sgro, 2011).

The highly variable life history of PFM on different host plants may reflect different host-associated biotypes as supported by an analysis of esterase isozyme patterns (Hua & Hua, 1995) and random amplified polymorphic DNA (RAPD) (Xu & Hua, 2004), although this

This article is protected by copyright. All rights reserved
is not yet been confirmed by direct studies on population differentiation in PFM (Kwon et al., 2017; Wang et al., 2015). Previous studies have shown that detoxification and chemosensory genes are associated with host adaptation of insects (Heckel, 2018; Rane et al., 2019; Wan et al., 2019). Understanding the genomic features of PFM may provide useful points to investigate host plant adaptation and reveal the genetic basis related to the potential invasiveness of PFM as well as other fruit boring pests (Kirk et al., 2013).

Except for genetic variations, the insect can adapt to the varied environment by plasticity, which often involves the expression change of related genes (Sgro et al., 2016). Transcriptome analysis provides an efficient way to identify the differentially expressed genes. Although the transcripts can be assembled de novo, a reference genome-guided analysis may help to make the analysis more precise.

Well-assembled genomes are increasingly becoming available as resources for tracing evolutionary adaptation across the Lepidoptera. Already there are substantial genomic resources for many moths (Chen et al., 2019; Cheng et al., 2017; Kanost et al., 2016; Lange et al., 2018; Ma et al., 2020; Pearce et al., 2017; Wan et al., 2019; Xia et al., 2004; Xiang et al., 2018; Xiao et al., 2020; You et al., 2013; Zhang et al., 2020) and butterflies (Ahola et al., 2014; Cong et al., 2015; Dasmahapatra et al., 2012; Lu et al., 2019; Nishikawa et al., 2015; Zhan et al., 2011) which are being used in comparative analyses to link genomic changes to phenotypes like the detoxification of compounds encountered in hosts (Rane et al., 2019). Available genomes provide abundant reference points for investigating evolution across the Lepidoptera, although most species sequences so far are from the Papilionoidea, Noctuoidea, Bombycoidea and Pyraloidea, with less genomic information available for the Carposinidae (Copromorphoidea) despite the importance of this group as agricultural pests.

In the present study, we report on a chromosome-level genome of PFM, which was de novo assembled based on sequences obtained from the PacBio and Illumina platforms and assembled at the chromosome level with the Hi-C technique. We compare features of the PFM genome with those of eleven other moths, focusing particularly on detoxification and chemosensory gene families important in host adaptation and pesticide resistance,
contributing to ecological niches occupied by species (Rane et al 2019). As an initial study using the newly assembled genome, we investigate transcriptomic changes induced by long-day and short-day photoperiods that induce diapause in PFM larvae, and we identify genes involved in these responses which are critical to climatic adaptation by PFM.

Materials and methods

Sample collection and rearing

We established a laboratory strain of PFM from 30 larvae collected from an apple orchard in the Beijing area of China in July 2018. This strain was maintained for five generations on apple (Malus pumila Mill) in the laboratory under 25 ± 1 °C, a relative humidity of 75 ± 5%, and a photoperiod of 15L : 9D (ND, normal-day condition). Eggs were laid on filter paper and moved to ripe apples before hatching. Larvae developed in the apples and the last (fifth) instar larvae left the fruit to pupate on prepared sawdust. Samples used in genome sequencing and RNA-seq were from this strain.

In order to induce diapause in larvae, we moved batches of newly hatched larvae to long-day and short-day photoperiodic conditions before they bored into apple for feeding. The long-day condition (LD) was set to a photoperiod of 22 L : 2 D, while the short-day condition (SD) was set to a photoperiod of 8 L : 16 D (Hua et al., 1998; Huang et al., 1976). Both treatments were conducted under 25 ± 1 °C, and a relative humidity of 75 ± 5%. The last instar larvae leaving the fruit were collected and stored in RNAlater at −80 °C (Sigma-Aldrich, St. Louis, USA) for subsequent RNA-seq library construction.

Genome sequencing

We extracted genomic DNA from 12 male pupae using MagAttract HMW DNA kit (Qiagen, Hilden, Germany) for Illumina library and PacBio library. The paired-end Illumina library with insert sizes of about 500 bp, was constructed using VAHTS™ Universal DNA Library Prep Kit for Illumina® V2 (Vazyme, Nanning, China) and sequenced on an Illumina Novaseq platform to obtain 150-bp paired-end reads. The raw reads generated were filtered by the software Trimmomatic v0.38 (Bolger et al., 2014). After filtering, we obtained 31.02 Gb of short clean reads.
reads (coverage: 77.24X). The sequencing data was used to estimate the genome size, heterozygosity and rate of duplication and polish de novo assemblies.

For long-read sequencing, SMRTbell libraries were constructed with Sequel® Sequencing Kit 3.0 (Pacific Biosciences, Menlo Park, CA, USA). Long DNA fragments of approximately 20 kb were sequenced on a PacBio Sequel sequencer (Pacific Biosciences, Menlo Park, CA, USA). Four SMRT cells were processed and 55.52 Gb subreads (mean subread length: 18.13 kb, subread N50 length: 32.84 kb, coverage: 138.2X) were obtained for contig-level genome assembly.

To assist the chromosome-level assembly, we used the Hi-C (High-throughput chromosome conformation capture) technique to capture genome-wide chromatin interactions (Belaghzal et al., 2017). Twenty 5th instar larvae were ground in 2% formaldehyde for cross-linking of cellular protein. Chromatin was digested with restriction enzyme MboI overnight. Then, the DNA ends were flatted, marked with biotin-14-dCTP and ligated with bridge linker. The samples were digested with proteinase K and purified by phenol-chloroform extraction. Biotins on unligated DNA fragments ends were removed with T4 DNA polymerase. Fragments were sheared into 200-600 base pairs using an S220 Focused-ultrasonicator (Covaris, U.S.). Biotin marked DNA fragments were enriched using streptavidin C1 magnetic beads. Illumina library was constructed from the enriched fragments using VAHTSTM Universal DNA Library Prep Kit for Illumina® V2 (Vazyme, Nanning, China) and sequenced on an Illumina Novaseq platform to obtain 150-bp paired-end reads. After removing the low-quality reads, 1,509 million clean reads were retained (coverage: 559.3X).

Genome survey

We used the k-mer method to survey the genome features of the PFM. The k-mer count histogram was obtained from Illumina paired-end sequencing data using Jellyfish v2.99 (Marçais & Kingsford, 2011) with 17, 21, 25 and 35 mers. Genome size, heterozygosity and rate of duplication were estimated by GenomeScope v1.0 (Vurture et al., 2017).

Genome assembly and evaluation

This article is protected by copyright. All rights reserved
Long reads generated from PacBio sequencing were corrected and assembled using CANU version 1.8 (Koren et al., 2017) with default parameters. The initial assembly was polished using Pilon v1.22 (Walker et al., 2014) with short reads from Illumina paired-end sequencing for three times. Two haplotypes in part of the genome might be assembled as separate primary contigs due to the high degree of heterozygosity (Roach et al., 2018). To correct these possible allelic contigs, we reassigned the polished assembly using the pipeline Purge Haplotigs to identify pairs of contigs that are syntenic and removed one of them (Roach et al., 2018), resulting in a contig-level genome.

Clean reads sequenced from the Hi-C library were aligned to the contig-level genome with an end-to-end algorithm implemented in Bowtie v2.3.5 (Langmead & Salzberg, 2012) according to the HiC-Pro strategy (Langmead & Salzberg, 2012; Servant et al., 2015). The Juicer v1.5 and 3D de novo assembly (3D-DNA) pipelines were used to assemble the contigs into a chromosome-level genome (Dudchenko et al., 2017; Durand et al., 2016). The completeness of the genome was evaluated through the analysis of single-copy orthologs (Simao et al., 2015), implemented in Benchmarking Universal Single-Copy Orthologs (BUSCO) v3.0.2 (Simao et al., 2015), based on the insecta_odb9 database (1,658 genes). Synteny between PFM and *Cydia pomonella* (Lepidoptera: Tortricidae) (Assembly accession: GCA_003425675.2) (Wan et al., 2019) and *Spodoptera litura* (Assembly accession: GCF_002706865.1) (Cheng et al., 2017) genomes were analyzed using TBtools v0.58 (Chen et al., 2020).

**Transcriptome sequencing and assembly**

To provide evidence of transcripts for genome structure annotation, we conducted RNA-seq for four developmental stages of egg, larva, pupae, and adults (male and female) reared under normal conditions as described above. To identify the differentially expressed genes between normal (ND) and diapausing larvae, we constructed another two RNA-seq libraries for the long-day (LD) and short-day (SD) induced 5th instar larvae. In total, 7 RNA-seq libraries were constructed, including one for eggs, three for larvae, one for pupae, one for male adults and one for female adults of PFM. One individual collected from each
stage/treatment was used (except for eggs, for which about 100 eggs were used) for RNA-seq and one library was constructed for each treatment. All libraries were prepared using VAHTSTM mRNA-seq V2 Library Prep Kit for Illumina according to the manufacturer’s instructions (Vazyme, NanJing, China) and sequenced on an Illumina Novaseq platform to obtain 150-bp paired-end reads. After removing the low quality reads with Trimmomatic v0.38 (Bolger et al., 2014), the reads were mapped to the chromosome-level genome using Hisat v2.2.0 (Kim et al., 2019) and assembled with StringTie v2.1.2 (Pertea et al., 2015). FPKM (Fragments Per Kilobase per Million) values of each annotated gene in each RNA-seq were estimated with cufflinks v2.2.1 (Kim et al., 2013).

We analyzed the differentially expressed gene (DEG) among the 5th instar larvae reared at different photoperiods (SD, ND and LD). The ND was used as a control, while the SD and LD were compared to the ND. Due to the lack of biological replicate, we used cufflinks v2.2.1 (Kim et al., 2013) in the assessment of DEGs, which allowed one repeat in the analysis. Because we analyzed the DEGs for the 5th instar larvae, which is the last developmental stage before diapause, we assumed that the fold change of DEGs related to diapause should be large. Thus, we used a fold-change $\geq 2$ and q-value $\leq 0.05$ as the cutoffs of significant genes between samples. For significantly expressed genes, up-regulated or down-regulated genes in both comparisons (ND vs. LD and ND vs. SD) were considered as genes related to diapause, while the FPKM values specifically high in the SD or LD condition were considered as light-induced genes. Gene expression visualization of DEGs was conducted with the Pheatmap R package.

Repeat element and non-coding RNA annotation

Repeats and transposable element families in the PFM genome were first detected by RepeatMasker v4.0.7 pipeline (Tarailo-Graovac & Chen, 2009) against the Insecta repeats within RepBase Update (http://www.girinst.org) and Dfam database (20170127), with RMBlast v2.10.0 as a search engine. The noncoding RNAs (ncRNA) were annotated by aligning the genomic sequence against RFAM v14.2 (http://rfam.xfam.org/) with BLASTN.
The tRNAs and rRNAs were predicted by tRNAscan-SE and RNAmmer (Lagesen et al., 2007; Lowe & Eddy, 1997).

**Protein-coding gene annotation and filtering**

We annotated protein-coding genes using *ab initio*, RNA-seq-based, and homolog-based methods in the MAKER v2.31.10 genome annotation pipeline (Cantarel et al., 2008). Augustus v3.2.3 (Stanke & Waack, 2003) and SNAP v2013-02-16 (Korf, 2004) were used for the *ab initio* gene prediction. For Augustus, we used the retrained parameters obtained in the above BUSCO analysis of genome assembly by invoking the Augustus retraining option. In the first round of annotation, we ran MAKER by providing transcriptome assemblies of PFM, protein sequences from eight lepidopteran species (*Bombyx mori*, *Trichoplusia ni*, *Ostrinia furnacalis*, *Bombyx mandarina*, *Galleria mellonella*, *Spodoptera litura*, *Helicoverpa armigera*, *Plutella xylostella*) and the Augustus model as evidence. The GFF3 file of first round annotation was used to train parameters of SNAP. In the next three rounds of annotation, GFF3 from the last round, Augustus and SNAP models were used as evidence.

The annotation results from the MAKER pipeline were filtered by using gene expression evidence, functional annotation results and Annotation Edit Distance (AED) value. Genes that had a FPKM value great than 0 in any RNA-seq were considered as real genes and retained in further analysis. Functional domains for proteins were identified using InterproScan 5.34-74.0 (Jones et al., 2014) against Pfam database v32.0 (El-Gebali et al., 2019). The gene models were filtered based on domain content and evidence support following Campbell et al. (2014). Finally, the annotations with AED < 0.75 were removed (Campbell et al., 2014).

Functions of the protein-coding genes and their GO (Gene Ontology) and KEGG (Kyoto Encyclopedia of Genes and Genomes) items were annotated using the software eggNOG-Mapper v1.0.3 (Huerta-Cepas et al., 2017), a tool for fast functional annotation of novel sequences using precomputed eggNOG-based orthology assignments, against the database EggNOG v5.0 (Huerta-Cepas et al., 2019).

**Orthology identification and phylogenetic inference**
Protein-coding genes from another 11 species of Lepidoptera as well as two species of Coleoptera and two species of Diptera were obtained from the NCBI genomes database for comparative analysis (Table 1). Orthologs were identified using OrthoFinder version 2.2.7 (Emms & Kelly, 2015) under default parameters. The phylogenetic tree was inferred in the OrthoFinder pipeline with an approximately-maximum-likelihood method implemented in FastTree v2.1.10 (Price et al., 2009) based on a concatenated multiple sequence alignment (MSA) of single-copy genes. The most likely category for each site was set using a Bayesian approach with a gamma prior. Amino acid sequences were aligned in MAFFT v7.450 (Katoh & Standley, 2013) with the G-INS-I algorithm.

**Manual annotation of circadian genes**

We further manually annotated well-studied circadian genes: period (PER), timeless (TIM), Clock (CLK), cycle (CYC) and cryptochrome (CRY), using BLAST v2.2.31 (Altschul et al., 1990). Reference protein sequences of insect circadian genes were obtained from the Uniprot database. Conserved domains within proteins were annotated against the conserved domain database (Lu et al., 2020). Circadian genes of the other 15 insect species were annotated in the same way. For a common domain of three genes (CLK, PER and CYC), a neighbor-joining tree was constructed using MEGA7 (Kumar et al., 2016) with 500 bootstrap replicates.

**Manual annotation of detoxification and chemosensory gene families**

We manually annotated five detoxification gene families and four chemosensory gene families, including cytochrome P450 monoxygenase (P450s), glutathione S-transferase (GSTs), carboxyl/cholinesterases (CCEs), uridine diphosphate (UDP)-glycosyltransferases (UGTs), ATP-binding cassette (ABC) transporters, gustatory receptors (GR), ionotropic receptors (IR), odorant-binding proteins (OBP), olfactory receptors (OR) genes. We used the bioinformatic pipeline BITACORA (Vizueta et al., 2019) to conduct HMMER v3.3 (Finn et al., 2011) and BLAST v2.2.31 (Altschul et al., 1990) analyses under a full mode. Hits were filtered with a default cutoff E-value of 10e-5. The HMMs of P450 were downloaded from Pfam v32.0 (El-Gebali et al., 2018), while other HMMs of detoxification gene families were...
created by HMMER v3.3 (Finn et al., 2011). Orthologs from *Bombyx mori* and *D. melanogaster* were used as evidence. The annotated genes were further filtered manually based on gene length and the presence of conserved domains. The longest transcripts of alternative splicing was kept. Genes with a length shorter than 80 amino acids were removed. Orthologs were aligned with the G-INS-I algorithm implemented in MAFFT v7.450 (Katoh & Standley, 2013). A neighbor-joining tree was constructed for each gene family using MEGA7 (Kumar et al., 2016) with 500 bootstrap replicates.

**Results and discussion**

*Features of the assembled genome*

The genome size of PFM is estimated to be 338.52-352.59 Mb through k-mer analysis depend on the k-mers used (k = 17, 21, 25, 35). The k-mer distributions showed double peaks, indicating that this genome has a high rate of duplication and heterozygosity. The estimated heterozygosity ranges from 1.06% to 1.15% and the rate of duplication range from 1.95% to 2.06% (Fig. 2a). The high heterozygosity of the PFM genome might be caused by the pooling of the samples for short-read sequencing. The heterozygosity of PFM is similar to the *Thrips palmi* (1.01% to 1.32%) (Guo et al., 2020) and the caddisfly *Stenopsyche tienmushanensis* (1.05%-1.10%) (Luo et al., 2018), which were also sequenced from pooling DNA of multiple individuals. The heterozygosity of PFM is higher than beet armyworm *Spodoptera exigua* (0.59%) (Zhang et al., 2019), the invading fall webworm (0.75% and 0.83%) (Wu et al., 2018) and *Cydia pomonella* (0.6%) (Wan et al., 2019), which were estimated from one individual or multiple individuals of sibling mating lines.

At the contig level, we assembled the PFM genome into 404.83 Mb sequences, including 275 contigs, with a contig N50 length of 2.62 Mb. Based on contig interaction frequency calculated from the pairs aligned to the contigs, the 275 contigs were clustered into 31 linkage groups (Fig. 2b). The longest contig group was 19.1 Mb while the shortest one was 6.30 Mb, with an N50 of 14.39 Mb. BUSCO analysis showed that 98.3% (single-copied gene: 97.4%, duplicated gene: 0.9%) of 1,658 genes were identified as
complete, 0.40% of genes were fragmented, while 1.3% of genes were missing in the assembled genome. The genome comprised 36.88% GC base pairs.

Synteny analysis showed that the PFM, *S. litura* and *C. pomonella* genome have a highly conserved gene order (Fig. 2c). Chromosome 01 of PFM was identified as the Z chromosome based on its high synteny to the Z chromosome of *S. litura* (Fig. 2c). Chromosome 04 of PFM shows high synteny to chromosome 02 of *S. litura*, which fused with Z chromosome in *C. pomonella*, forming the neo-Z chromosome (chromosome 01) (Wan et al., 2019). In conclusion, PFM has similar chromosomes as *S. litura*, including 30 autosomes and a Z chromosome. The female-specific W chromosome of PFM was not determined in this genome, since male pupae were used for sequencing. Identification of W chromosome is challenged by its high degenerate, being gene-poor and repeat-rich in Lepidoptera (Bergero & Charlesworth, 2009; Wan et al., 2019). The chromosome-level assembly of the PFM genome provides resources for understanding chromosome evolution in the Lepidoptera (Ahola et al., 2014).

**Genome annotation**

We identified 27,598 protein-coding genes in the 1st round of MAKER annotation. BUSCO analysis revealed 91.5% of the evaluated single-copy genes were identified as complete. After three rounds of MAKER annotation, the number of genes increased to 50,676, while the proportion of complete single-copy genes was up to 95.0%. After filtering based on gene expression analysis, functional domains and AED values, 21,695 genes remained. BUSCO analysis showed that 94.8% (single-copied gene: 90.5%, duplicated gene: 4.3%) of the evaluated single-copy genes were identified as complete, 3.6% of the genes were fragmented, and 1.6% of the genes were missing in the annotated gene set. In total, 18,397 genes (84.80%) were functionally annotated, of which 5,303 (24.44%) and 3,829 (17.65%) genes were annotated to GO terms and KEGG KOs respectively. We predicted 53 rRNAs, 11,076 tRNAs, 20 small nuclear RNAs, and 48 micro RNAs in the PFM genome based on Rfam databases.

In total, 35.70 Mb (8.95%) of the genome was identified to be repeat DNA. Overall, 179,142 transposable elements (TEs, 35.14 Mb), including 139,294 retroelements (16,818 short interspersed nuclear elements (SINEs), 107,497 long interspersed nuclear elements

This article is protected by copyright. All rights reserved
(LINEs) and 14,979 long terminal repeats (LTR)) and 39,848 DNA transposons were identified
(Table 1, Table S1). The PFM genome shows the lowest repeat DNA among seven
chromosome-level assembled genomes from the Lepidoptera (Table 1).

Orthology and phylogenetic relationships of lepidopterans
OrthoFinder assigned 320,821 genes (93.41% of total) to 15,076 orthogroups for the 16
species compared. Fifty percent of the assigned genes were in orthogroups with 28 or more
genes (G50 was 28) and were contained in the largest 3,174 orthogroups (O50 was 3,174).

There were 947 single-copy genes with 364,262 reliable sites retained for phylogenetic
inference. The topology is congruent with previously inferred phylogenetic relationships of
Lepidoptera, in which no representative of the Copromorphoidea was included (Wan et al.,
2019). Current molecular phylogenetic studies have not resolved the phylogenetic
relationship between Copromorphoidea and Papilionoidea (Mitter et al., 2017). Our result
supports the notion that PFM from the Copromorphoidea forms a sister-group relationship
to the butterfly D. plexippus (Papilionoidea), rather than a sister group between
Copromorphoidea/Papilionoidea and Pyraloidea + (Noctuoidea + Bombycoidea) (Fig. 3).

We investigated orthogroups shared by PFM and four species of Lepidoptera,
representing different clades of the phylogenetic tree of Lepidoptera (Fig. S1). There were
7,827 orthogroups (60.5% of 12,938 orthogroups) shared by all five lepidopteran species
and 1,549 orthogroups shared by four species except for C. pomonella. We identified 357
orthogroups specific to PFM, fewer than that of B. mori (406), but higher than other three
lepidopteran species (Fig. S1).

Evolution of circadian genes
Five core genes of the circadian clock were annotated in the PFM genome and the other
reference insect species. The PER gene was not found in currently assembled genomes of
Cydia pomonella and Anoplophora glabripennis (Fig. 4). Two types of CRY gene were
annotated in 16 species, mammalian-type cryptochrome (CRY-m) and Drosophila type
cryptochrome (CRY-d). CRY-d is a UV- and blue-light photoreceptor and is not considered as
a core component of the circadian clock in D. melanogaster (Goto, 2013). In circadian clock
model established from Danaus plexippus (L.), CRY-m function as a transcriptional repressor
instead as a photoreceptor (Yuan et al., 2007; Zhu et al., 2008), and it was incorporated into
the core component of the circadian clock as a negative regulator, like PER and TIM. For
most of the 16 insects, two types of CRY genes were found, while only CRY-m was found in two Coleoptera species, and only CRY-d was found in *D. melanogaster*. In the wasp *Nasonia vitripennis*, only CRY-m was found and TIM does not exist in the genome (Schurko et al., 2010). In the honey bee *Apis merifella*, TIM was lost (Rubin et al., 2006). In the PFM, FPKM values of CRY-m were higher than CRY-d in each stage, indicating that CRY-m may be a major element in the circadian clock of PFM. These considerable differences in the gene composition of the circadian clock indicate a diverse circadian clock model among insects.

Domains of circadian genes were conserved among the 16 species (Fig. 4). The PAS domains are widely present in photoreceptors and circadian proteins of many eukaryotes (Taylor & Zhulin, 1999). For insects investigated in this study, CLK, CYC and PER genes have two PAS domains. In previous studies, the two PAS domains were defined as PAS-A and PAS-B. In this study, PAS-A is equivalent to PAS and PAS-fold (Fig. 4 and Fig. 5). Besides PAS domains, CLK and CYC genes have other common domains bHLH before two PAS domains (Fig. 4). Both CRY-d and CRY-m gene have a PhrB domain, and TIM gene has a TIMELESS domain (Fig. 4). We constructed phylogenetic relationships among the PAS domains for the 16 species. The phylogenetic tree of PAS domains revealed six clades, corresponding to two PAS domains of three genes (CLK, CYC and PER) (Fig. 5).

**Gene expression in diapause and non-diapause PFM**

Compared with larvae that developed under a normal day photoperiod, 11 genes were significantly up-regulated and 9 genes were down-regulated in larvae that developed under long-day or short-day photoperiods (Table S2, Fig. S2). Genes highly expressed in pre-diapause larvae (SD and LD photoperiod) included genes encoding CUSOD2 (CS_07203), an enzyme that destroys radicals, and that plays an important role in diapause and cold tolerance of insect (Bi et al., 2014; He et al., 2013; Isobe et al., 2006; Kim et al., 2010; Sim & Denlinger, 2011; Zhao & Shi, 2009). We identified a cytochrome P450 gene (CS_20496) showing weak expression in pre-diapause larvae and high expression in the other stages, which was also found in diapausing larvae of the wild silk moth, *Antheraea yamamai* (Yang et al., 2008).

We identified 44 genes specifically up-regulated under a long-day photoperiod, and 14 genes specifically up-regulated under a short-day photoperiod (Table S2, Fig. S2). Four
genes (CS_04235, CS_05017, CS_15183, CS_01854) related to digestion of proteins were
up-regulated in larvae developing under a long-day photoperiod. This is congruent with
previous reports suggesting that the photoperiod had significant effects on digestive
enzyme activity (Espinosa-Chaurand et al., 2017; Ramzanzadeh et al., 2016; Shan et al., 2008;
Subala & Shivakumar, 2017). The functional link of many of these genes to diapause is not
really clear. The circadian genes, which are important in diapause in another moth (Kozak et
al., 2019), did not show significant changes for larvae under different photoperiods.

Evolution of detoxification and chemoreceptor genes

We manually identified 95 P450s, 76 GSTs, 63 CCEs, 27 UGTs, and 93 ABCs in the PFM
genome (Table 2; Fig. S3). PFM had the lowest number of UGT genes and CCE genes, along
with two other moths located at basal lineages of the Lepidoptera. The number of P450
genes in PFM are the second lowest, only slightly higher than Bombyx mandarina. The
number of ABCs genes in PFM are at an intermediate level. We found that the PFM had the
highest number of GST genes. These results suggest that PFM may have a unique way of
detoxication with reduced importance of UGT and P450 when compared to the other moths.
This may have implications for pesticide responses in PFM give that these detoxification
genes can respond in different ways to various pesticides in moths (Hu et al., 2019).

The chemosensory system plays an important role in locating food, shelter, mates, and
oviposition sites (Wan et al., 2019). We manually annotated several well-studied gene
families of the chemosensory system; we identified 41 GRs, 49 IRs, 56 OBPs, and 176 ORs in
the PFM genome (Table 2; Fig. S4). PFM had the lowest number of GR genes and the
second-lowest number of IR genes, OBP genes, and OR genes following the C. pomonella
when compared to the other lepidopterans. Both PFM and C. pomonella have a lower
number of chemoreceptor genes. The PFM and C. pomonella mainly feed on apple and pear,
with a narrow host range. Females of these two moths lay eggs on the surface of their host
fruits, and the hatched larvae bore into the fruit directly without long-time searching
behavior (van der Geest & Evenhuis, 1991). The narrow host range and feeding behavior of
larva may explain the lower number of chemoreceptor genes in their genomes.

Conclusions
We assembled the chromosome-level genome for the PFM using PacBio long-read and Hi-C technology. This is the first assembled genome for the superfamily Copromorphoidea. This novel genomic resource allowed us to explore possible genes in PFM associated with adaptation to environmental factors. We identified five core genes relating to the circadian rhythm in PFM and annotated models for each gene. Using the genome as a reference, we investigated DEGs related to the diapause of OFM, which may point to candidate genes. Given the expression of long-day and short-day diapause by PFM, this moth species will be a useful model to further investigate adaptive shifts involving diapause, particularly by combining genomic information with intraspecific comparisons across geographic gradients (Ragland et al., 2019). The assembled genome provides a resource for further comparative studies of moths and butterflies, particularly with respect to life cycle evolution and parallel evolution in detoxification and chemosensory functions.

**Data Availability Statement**

The Whole Genome assembly has been deposited in the Genome repository of NCBI (accession numbers: CP053148-CP053178) under BioProject PRJNA627116. Raw reads obtained for genome assembly were deposited in the Sequence Read Archive (SRA) repository (accession numbers: SRR12328811 and SRR12336732). Scripts used for assembling and annotating the genome and sequences of manually annotated gene families were deposited in the Dryad repository (https://doi.org/10.5061/dryad.m0cfxpp1j).

**Author contributions**

Shu-Jun Wei conceived and designed the study; Jin-Cui Chen and Ya-Jun Gong conducted the collection and rearing of the insect; Li-Jun Cao conducted the molecular works; Li-Jun Cao, Wei Song, Lei-Yue, Shao-Kun Guo, and Shu-Jun Wei analyzed the data; Li-Jun Cao, Shu-Jun Wei, and Ary Hoffmann discussed the results and wrote the manuscript.

**Acknowledgments**

We thank Qiang Gao for his help on the assembly of Hi-C data and Qiang Gong for the assistance of insect rearing. This research was supported by the National Natural Science Foundation of China (31901884, 32070464), the Natural Science Foundation of Beijing Municipality (6184037), Joint Laboratory of Pest Control Research Between China and
Australia (Beijing Municipal Science & Technology Commission), and the Beijing Key
Laboratory of Environmentally Friendly Pest Management on Northern Fruits (BZ0432).

Reference

(2014). The Glanville fritillary genome retains an ancient karyotype and reveals
doi:10.1038/ncomms5737

Altschul, S., Gish, W., Miller, W., Myers, E., & Lipman, D. (1990). Basic local alignment search

high-resolution genome-wide mapping of chromosome conformation. Methods, 123,
56-65.

sequence data. Bioinformatics, 30(15), 2114-2120.
doi:10.1093/bioinformatics/btu170

Using MAKER and MAKER-P. Current Protocols in Bioinformatics, 48(1),
4.11.11-4.11.39. doi:10.1002/0471250953.bi0411s48

MAKER: an easy-to-use annotation pipeline designed for emerging model organism
genomes. Genome Research, 18(1), 188-196. doi:10.1101/gr.6743907

This article is protected by copyright. All rights reserved


This article is protected by copyright. All rights reserved


He, C., Meng, Q.-K., Yang, X.-B., & Hua, L. (2013). Carbohydrate metabolism and antioxidant defense during diapause development in larvae of oriental fruit moth (*Grapholita...*
molestia) at low temperature. *International Journal of Agriculture and Biology, 15*(1), 101-106.


This article is protected by copyright. All rights reserved.


This article is protected by copyright. All rights reserved


This article is protected by copyright. All rights reserved


This article is protected by copyright. All rights reserved


This article is protected by copyright. All rights reserved
729 identification and annotation of gene families in genome assemblies. bioRxiv,
730 593889. doi:10.1101/593889
735 (2014). Pilon: an integrated tool for comprehensive microbial variant detection and
736 genome assembly improvement. PLoS ONE, 9(11), e112963.
737 doi:10.1371/journal.pone.0112963
739 chromosome-level genome assembly of Cydia pomonella provides insights into
740 chemical ecology and insecticide resistance. Nature Communications, 10,
741 doi.org/10.1038/s41467-41019-12175-41469. doi:ARTN 4237
742 10.1038/s41467-019-12175-9
744 Carposinidae) in its native range consists of two sympatric cryptic lineages as
746 doi:10.1093/jisesa/iev063
748 Patterns of genetic variation among geographic and host-plant associated
749 populations of the peach fruit moth Carposina sasakii (Lepidoptera: Carposinidae).
750 BMC Evolutionary Biology, 17(1), 265. doi:10.1186/s12862-017-1116-7
752 genomes yield insights into rapid adaptation of invasive species. Nature Ecology and
753 Evolution, 3(1), 105-115. doi:10.1038/s41559-018-0746-5

This article is protected by copyright. All rights reserved


This article is protected by copyright. All rights reserved


This article is protected by copyright. All rights reserved
### Table 1 Features of chromosome-level genomes in the Lepidoptera

<table>
<thead>
<tr>
<th>Features</th>
<th>Csas</th>
<th>Cpom</th>
<th>Tni</th>
<th>Bmor</th>
<th>Slit</th>
<th>Mcin</th>
<th>Hmel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome size (Mb)</td>
<td>399.04</td>
<td>772.89</td>
<td>368.2</td>
<td>431.7</td>
<td>438.32</td>
<td>393</td>
<td>269</td>
</tr>
<tr>
<td>Karyotype</td>
<td>2n=64</td>
<td>2n=56</td>
<td>2n=54</td>
<td>2n=56</td>
<td>2n=62</td>
<td>2n=62</td>
<td>2n=42</td>
</tr>
<tr>
<td>No. contigs</td>
<td>275</td>
<td>2221</td>
<td>26,605</td>
<td>15,018</td>
<td>13,636</td>
<td>49,851</td>
<td>NA</td>
</tr>
<tr>
<td>No. scaffolds</td>
<td>NA</td>
<td>1717</td>
<td>6181</td>
<td>7397</td>
<td>3597</td>
<td>8262</td>
<td>3807</td>
</tr>
<tr>
<td>No. CHR*</td>
<td>31A+Z+W</td>
<td>27A+Z+W</td>
<td>26A+Z+W</td>
<td>27A+Z</td>
<td>30A+Z</td>
<td>30A+Z</td>
<td>20A+Z</td>
</tr>
<tr>
<td>Contig N50 (kb)</td>
<td>2620</td>
<td>862.49</td>
<td>621.9</td>
<td>15.5</td>
<td>68.35</td>
<td>13</td>
<td>51</td>
</tr>
<tr>
<td>Scaffold N50 (Mb)</td>
<td>NA</td>
<td>8.92</td>
<td>14.2</td>
<td>3.7</td>
<td>0.915</td>
<td>0.119</td>
<td>0.277</td>
</tr>
<tr>
<td>BUSCO genes (%)</td>
<td>98.20%</td>
<td>98.5</td>
<td>97.8</td>
<td>97.7</td>
<td>98.3</td>
<td>91.5</td>
<td>97.4</td>
</tr>
<tr>
<td>Repeat (%)</td>
<td>11.33</td>
<td>42.87</td>
<td>20.5</td>
<td>43.6</td>
<td>31.83</td>
<td>28</td>
<td>24.94</td>
</tr>
<tr>
<td>G+C (%)</td>
<td>36.96</td>
<td>37.43</td>
<td>35.6</td>
<td>37.3</td>
<td>36.5</td>
<td>33.0</td>
<td>NA</td>
</tr>
<tr>
<td>No. genes</td>
<td>23,227</td>
<td>17,184</td>
<td>14,043</td>
<td>14,623</td>
<td>15,317</td>
<td>16,667</td>
<td>12,669</td>
</tr>
</tbody>
</table>

Table 2 Number of genes in five detoxification families across species of Lepidoptera

<table>
<thead>
<tr>
<th>Species</th>
<th>P450</th>
<th>GST</th>
<th>CCE</th>
<th>ABC</th>
<th>UGT</th>
<th>GR</th>
<th>IR</th>
<th>OBP</th>
<th>OR</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pxyl</td>
<td>163</td>
<td>55</td>
<td>85</td>
<td>219</td>
<td>38</td>
<td>68</td>
<td>61</td>
<td>91</td>
<td>225</td>
<td>You et al. (2013)</td>
</tr>
<tr>
<td>Csas</td>
<td>95</td>
<td>76</td>
<td>63</td>
<td>93</td>
<td>27</td>
<td>41</td>
<td>49</td>
<td>56</td>
<td>176</td>
<td>This study</td>
</tr>
<tr>
<td>Cpom</td>
<td>136**</td>
<td>30**</td>
<td>73**</td>
<td>47**</td>
<td>30**</td>
<td>65**</td>
<td>39**</td>
<td>50**</td>
<td>85**</td>
<td>Wan et al. (2019)</td>
</tr>
<tr>
<td>Dple</td>
<td>107</td>
<td>35</td>
<td>73</td>
<td>76</td>
<td>47</td>
<td>74</td>
<td>62</td>
<td>68</td>
<td>253</td>
<td>Zhan et al. (2011)</td>
</tr>
<tr>
<td>Tni</td>
<td>143</td>
<td>51</td>
<td>122</td>
<td>71</td>
<td>68</td>
<td>85</td>
<td>74</td>
<td>91</td>
<td>274</td>
<td>Chen et al. (2019)</td>
</tr>
<tr>
<td>Slit</td>
<td>182</td>
<td>66</td>
<td>153</td>
<td>223</td>
<td>64</td>
<td>107</td>
<td>62</td>
<td>87</td>
<td>261</td>
<td>Cheng et al. (2017)</td>
</tr>
<tr>
<td>Harm</td>
<td>122</td>
<td>57</td>
<td>105</td>
<td>76</td>
<td>54</td>
<td>120</td>
<td>71</td>
<td>79</td>
<td>257</td>
<td>Song et al. (2015)</td>
</tr>
<tr>
<td>Bman</td>
<td>94</td>
<td>37</td>
<td>94</td>
<td>64</td>
<td>48</td>
<td>69</td>
<td>56</td>
<td>74</td>
<td>218</td>
<td>Xiang et al. (2018)</td>
</tr>
<tr>
<td>Bmor</td>
<td>156</td>
<td>51</td>
<td>149</td>
<td>108</td>
<td>50</td>
<td>85</td>
<td>57</td>
<td>93</td>
<td>243</td>
<td>Xia et al. (2004)</td>
</tr>
<tr>
<td>Msex</td>
<td>164</td>
<td>66</td>
<td>137</td>
<td>103</td>
<td>51</td>
<td>89</td>
<td>76</td>
<td>79</td>
<td>261</td>
<td>Kanost et al. (2016)</td>
</tr>
<tr>
<td>Gmel</td>
<td>137</td>
<td>44</td>
<td>75</td>
<td>72</td>
<td>58</td>
<td>95</td>
<td>81</td>
<td>64</td>
<td>295</td>
<td>Lange et al. (2018)</td>
</tr>
<tr>
<td>Ofur</td>
<td>126</td>
<td>48</td>
<td>115</td>
<td>112</td>
<td>46</td>
<td>93</td>
<td>67</td>
<td>75</td>
<td>270</td>
<td>Ma et al. (2020)</td>
</tr>
</tbody>
</table>

glutathione S-transferase; CCEs, carboxyl/cholinesterases; UGT, uridine diphosphate-glycosyltransferases; ABC, ATP-binding cassette transporters; GR, gustatory receptors; IR, ionotropic receptors; OBP, odorant-binding proteins; OR, olfactory receptors.
Figure legends

**Fig. 1** Eggs (A), larva (B), cocoons (C) and adult (D) of the peach fruit moth *Carposina sasakii* (A-D) and the damage symptoms to apple (E-G). The hatched larva bores into apple usually near the calyx with white secreta near the boring hole (E); the damaged apple showing shrinkage (F); damage from larvae boring and developing in the apple (G).

**Fig. 2** Genome features of *Carposina sasakii*. (a) GenomeScope analysis of genome size, heterozygosity and duplicate rate using k-mers (K = 17) count histogram, indicating a genome size of 338.52 Mb, a heterozygosity of 1.06%, and a duplication rate of 2.06%; (b) Genome-wide all-by-all Hi-C interaction identified 31 linkage groups; (c) Synteny between *Carposina sasakii* (Csas) and *Cydia pomonella* (Cpom) and *Spodoptera litura* (Slit) genomes reveal highly conserved gene order and chromosomal fusion or split events in the three moths.

**Fig. 3** Phylogenetic tree of PFM with 15 insect genomes including 11 other Lepidoptera. The phylogeny was inferred from 947 single-copy genes with 364,262 reliable sites by an approximately-maximum-likelihood method. All nodes received bootstrap support of 100.

**Fig. 4** Schematic arrangement of the domains of five circadian genes including *period* (PER), *timeless* (TIM), *Clock* (CLK), *cycle* (CYC) and *cryptochrome* (CRY-m, CRY-d) in *Carposina sasakii* and other 15 insects. Boxes in different color show different domains, while numbers under boxes show the postion of domains on protein sequences. Species and their taxonomic status are shown on the left: Tcas, *Tribolium castaneum*; Agla, *Anoplophora glabripennis*; Agam, *Anopheles gambiae*; Dmel, *Drosophila melanogaster*; Pxy, *Plutella xylostella*; Cpom, *Cydia pomonella*; Csas, *Carposina sasakii*; Dple, *Danaus plexippus*; Tni, *Trichoplusia ni*; Slit, *Spodoptera litura*; Harm, *Helicoverpa armigera*; Bmor, *Bombyx mori*; Bman, *Bombyx mandarina*; Msex, *Manduca sexta*; Gmel, *Galleria mellonella*; Ofur, *Ostrinia furnacalis*.
Fig. 5 Phylogenetic relationships of two PAS domains in three circadian genes: period (PER), Clock (CLK) and cycle (CYC). Each tip is labeled by the name of domain, gene and species. Abbreviations of species are same as in Fig. 4. Six clades shaded in different color reveal two domains of three genes, while one domain of CLK gene has two different types among species. Tips in red show the position of Carposina sasakii.

Supplemental Information

Table S1 Classification of repeat sequences identified in the Carposina sasakii genome

Table S2 Gene expression of diapause and day length related genes across different developmental stages

Fig. S1 Orthogroups shared by five Lepidoptera species of Carposina sasakii, Cydia pomonella, Bombyx mori, Ostrinia furnacalis and Helicoverpa armigera.

Fig. S2 Gene expression of diapause and day length related genes across different developmental stages based on Carposina sasakii transcriptome data

Fig. S3 Phylogenetic relationships of Carposina sasakii ATP-binding cassette (ABC) transporter gene families (a), carboxyl/cholinesterases (CCEs) (b), glutathione S-transferase (GSTs) (c), cytochrome P450 (P450) (d) and UDP-glycosyltransferases (UGTs) (e) in comparison with Bombyx mori and Cydia pomonella.

Fig. S4 Phylogenetic relationships of Carposina sasakii gustatory receptors (GR) (a), ionotropic receptors (IR) (b), odorant-binding proteins (OBP) (c), olfactory receptors (OR) (d) in comparison with Bombyx mori
This article is protected by copyright. All rights reserved.
This article is protected by copyright. All rights reserved.
Author/s:
Cao, L-J; Song, W; Yue, L; Guo, S-K; Chen, J-C; Gong, Y-J; Hoffmann, AA; Wei, S-J

Title:
Chromosome-level genome of the peach fruit moth Carposina sasakii (Lepidoptera: Carposinidae) provides a resource for evolutionary studies on moths

Date:
2020-11-26

Citation:

Persistent Link:
http://hdl.handle.net/11343/276651