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[2], Shaokun Guo[1], Jin-Cui Chen[2], Ya-Jun Gong[1], Ary Hoffmann[3], and Shu-Jun Wei[2]

[1]Beijing Academy of Agriculture and Forestry Sciences
[2]Beijing Academy of Agricultural and Forestry Sciences
[3]The University of Melbourne Bio21 Molecular Science and Biotechnology Institute

August 14, 2020

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 32 linkage groups assisted by the Hi-C technique, including 30 autosomes, a female specific W chromosome and a Z chromosome. BUSCO analysis showed that 98.2% genes were complete and 0.4% of genes were fragmented, while 1.4% of genes were missing in the assembled genome. In total, 23,218 protein-coding genes were predicted, of which 82.72% 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 ABC and GST genes in PFM than in most other lepidopterans, in contrast to a lower number of UGT genes, 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.

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 (D. Kwon, Kwon, Kim, & Yang, 2018; Y. Z. Wang et al., 2017).

One possible of 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, Honma, Masaki, & Zoology, 1961). Both long-day and short-day photoperiods induce diapause in the last instar of PFM larvae, resulting in a diapausing cocoon (B.-Z. Hua, Zeng, & Zhang, 1998; Huang, Wang, Ye, Zhang, & Zhang, 1976). The life cycle of PFM can be univoltine or bivoltine, depending on photoperiods encountered and environmental factors like humidity (Chiba & Kobayashi, 1985; D.-S. Kim, Lee, & Yiem, 2000; Sato & Ishitani, 1976). Temperature also affects the occurrence of PFM through effects on development rate and the emergence of the overwintering generation from diapause (D. S. Kim, Lee, & Yiem, 2001; B. 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 populations of the PFM among geographical populations across its native range in China (Y. Z. Wang et al., 2017). 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 (L. Hua & Hua, 1995) and random amplified polymorphic DNA (RAPD) (Xu & Hua, 2004), although this is not yet been confirmed by direct studies on population differentiation in PFM (D. H. Kwon, Kim, Kim, Lee, & Yang, 2017; J. Wang et al., 2015).

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 (W. 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; S. Zhang et al., 2020) and butterflies (Ahola et al., 2014; Cong, Borek, Otwinowski, & Grishin, 2015; Dasmahapatra et al., 2012; Lu et al., 2019; Nishikawa et al., 2015; Zhan, Merlin, Boore, & Reppert, 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 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. 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 degC (Sigma-Aldrich, St. Louis, USA) for subsequent RNA-seq library construction.

Genome sequencing

We extracted genomic DNA from 12 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^{TM} Universal DNA Library Prep Kit for Illumina(r) 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, Lohse, & Usadel, 2014). After filtering, we obtained 31.02 Gb of short clean reads (coverage: 77.24X). The sequencing data was used to survey genome feature and polish *de novo* assemblies.

For long-read sequencing, SMRTbell libraries were constructed with Sequel(r) 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, Dekker, & Gibcus, 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 VAHTS^{TM} Universal DNA Library Prep Kit for Illumina(r) 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 (Marcais & 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

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, Schmidt, & Borneman, 2018). To corrected 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, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 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 (C. 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 5thinstar 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. 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 (D. Kim, Paggi, Park, Bennett, & Salzberg, 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 (D. Kim et al., 2013).

Differential gene expression among larvae reared at different photoperiod (SD, ND and LD) was assessed using cufflinks v2.2.1 (D. Kim et al., 2013). Genes with a fold-change [?] 2 and q-value [?] 0.05 were considered significant differentially expression genes (DEGs) 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 were 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 (S. El-Gebali et al., 2019). The gene models were filtered based on domain content and evidence support following Campbell, Holt, Moore, and Yandell (2014). Finally, the annotations with AED < 0.75 were removed (Campbell et al., 2014).

Functions of the protein-coding genes were annotated using the software eggNOG-Mapper v1.0.3 (Jaime 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 (J. 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, Dehal, & Arkin, 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, Gish, Miller, Myers, & Lipman, 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, Stecher, & Tamura, 2016) with 500 bootstrap replicates.

Manual annotation of detoxification gene families

We manually annotated five detoxification gene families of cytochrome P450 monooxygenase (P450s), glutathione S-transferase (GSTs), carboxyl/cholinesterases (CCEs), UDP-glycosyltransferases (UGTs) and ATP-binding cassette (ABC) transporters. We used the bioinformatic pipeline BITACORA (Vizueta, Sanchez-Gracia, & Rozas, 2019) to conduct HMMER v3.3 (Finn, Clements, & Eddy, 2011)and BLAST v2.2.31 (Altschul et al., 1990) analyses under a full mode. Hits were filtered with a default cut-off E-value of 10e-5. The HMMs of P450 were downloaded from Pfam v32.0 (Sara 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. 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 rate of duplication ranges from 1.95% to 2.06% (Fig. 2a).

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 32 linkage groups (Fig. 2b). The longest contig group was 19.1 Mb while the shortest one was 2.63 Mb, with an N50 of 14.39 Mb. BUSCO analysis showed that 98.2% (single-copied gene: 97.2%, duplicated gene: 1.0%) of 1,658 genes were identified as complete, 0.40% of genes were fragmented, while 1.4% of genes were missing in the assembled genome. The genome comprised 36.96% GC base pairs.

Synteny analysis showed that the PFM, *S. litura* and *C. pomonella* genome have a highly conserved gene order (Fig. 2c). PFM has similar chromosomes as *S. litura* , including 30 autosomes, a Z chromosome (Chr01) and a female specific W chromosome, while *C. pomonella* has undergone three fusion events, resulting in 27 autosomes, a W chromosome, and a neo-Z chromosome arising from a Z-autosomal fusion (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 29,228 protein-coding genes in the 1st round of MAKER annotation. BUSCO analysis revealed 91.9% of the evaluated single-copy genes were identified as complete. After three rounds of MAKER annotation, the number of genes increased to 52,667, while the proportion of complete single-copy genes was up to 95.2%. After filtering based on gene expression analysis, functional domains and AED values, 23,218 genes remained. BUSCO analysis showed that 95.0% (single-copied gene: 94.1%, duplicated gene: 1.1%) of the evaluated single-copy genes were identified as complete, 1.6% of the genes were fragmented, and 3.2% of the genes were missing in the annotated gene set. In total, 19,206 genes (82.72%) were functionally annotated, of which 5,970 (25.71%) and 3,134 (13.50%) genes 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, 45.5 Mb (11.33%) of the genome was identified to be repeat DNA. Overall, 259,729 transposable elements (TEs) including 125,601 retroelements (17,962 short interspersed nuclear elements (SINEs), 95,657 long interspersed nuclear elements (LINEs) and 11982 long terminal repeats (LTR)) and 34,478 DNA transposons were identified.

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, Davis, & Cummings, 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. 3a).

We investigated orthogroups shared by PFM and four species of Lepidoptera representing different clades of the phylogenetic tree of Lepidoptera (Fig 3b). 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. 3b).

Evolution of circadian genes

Five circadian genes 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* . Two types of CRY gene were annotated in 16 species, mammalian-type cryptochrome (CRY-m) and Drosophila type cryptochrome (CRY-d). For most of the 16 insects, two types of CRY gene were found, while only CRY-m was found in two Coleoptera species and only CRY-d was found in *Drosophila melanogaster* . 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. Domains of circadian genes were conserved among the 16 species (Fig. 4). PAS domains were common in CLK, CYC and PER genes. The phylogenetic tree of PAS domains revealed six clades, corresponding to two domains of three genes (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 S1, Fig. S1**). 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, Yang, Yu, Shu, & Zhang, 2014; He, Meng, Yang, & Hua, 2013; Isobe et al., 2006; Y. I. 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, Tanaka, Kuwano, & Suzuki, 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 S1, Fig. S1**). 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 photoperiod had significant effects on digestive enzyme activity (Espinosa-Chaurand, Vega-Villasante, Carrillo-Farnes, & Nolasco-Soria, 2017; Ramzanzadeh, Yeganeh, JaniKhalili, & Babaei, 2016; Shan, Xiao, Huang, & Dou, 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 genes

We manually identified 96 P450s, 77 GSTs, 63 CCEs, 28 UGTs, and 104 ABCs in the PFM genome (Table 2; Fig. S2). PFM had the highest number of GST genes and the second highest number of ABCs following the *O. furnacalis* when compared to the other lepidopterans. The number of P450 and CCE genes in PFM are at an intermediate level. We found that PFM had the lowest number of UGT genes, along with two other moths located at basal lineages of the Lepidoptera. These results suggest that PFM may have a unique way of detoxication with a reduced importance of UGT 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).

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 circadian rhythm in PFM and annotated models for each gene. Using the genome as a reference, we identified DEGs related to 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, Armbruster, & Meuti, 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 functions.

Data Availability Statement

The Whole Genome assembly has been deposited in the Genome repository of NCBI (accession numbers: CP053148-CP053179) under BioProject PRJNA627116 (reviewer link:*https://dataview.ncbi.nlm.nih.gov/object/PRJNA627116?reviewer=rvjh98ap96u9pqguc8k17dl0cb*). Raw reads obtained for genome assembly were deposited in the Sequence Read Archive (SRA) repository (accession numbers: SRR12328811 and SRR12336732). Gene sequences of manually annotated 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 National Natural Science Foundation of China (31901884), the Natural Science Foundation of Beijing Municipality (6184037), Beijing Academy of Agriculture and Forestry Sciences (BAAFS) - The University of Melbourne (UOM) Joint Laboratory on Pest Control Research, and the Beijing Key Laboratory of Environmentally Friendly Pest Management on Northern Fruits (BZ0432).

Reference

Ahola, V., Lehtonen, R., Somervuo, P., Salmela, L., Koskinen, P., Rastas, P., . . . Hanski, I. (2014). The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. *Nature Communications, 5* , 4737. doi:10.1038/ncomms5737

Altschul, S., Gish, W., Miller, W., Myers, E., & Lipman, D. (1990). Basic local alignment search tool. *Journal of molecular biology, 215* (3), 403-410.

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.

Bi, Z., Yang, X., Yu, W., Shu, J., & Zhang, Y. (2014). Diapause-Associated Protein3 Functions as Cu/Zn Superoxide Dismutase in the Chinese Oak Silkworm (*Antheraea pernyi*). *PLoS ONE, 9* (3). doi:10.1371/journal.pone.0090435

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

Campbell, M. S., Holt, C., Moore, B., & Yandell, M. (2014). Genome Annotation and Curation Using MAKER and MAKER-P. *Current Protocols in Bioinformatics, 48* (1), 4.11.11-14.11.39. doi:10.1002/0471250953.bi0411s48

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

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., & Xia, R. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Molecular Plant* , doi.org/10.1016/j.molp.2020.1006.1009. doi:10.1101/289660 %J bioRxiv

Chen, W., Yang, X., Tetreau, G., Song, X., Coutu, C., Hegedus, D., . . . Wang, P. (2019). A high-quality chromosome-level genome assembly of a generalist herbivore, *Trichoplusia ni* . *Molecular Ecology Resources, 19* (2), 485-496. doi:10.1111/1755-0998.12966

Cheng, T., Wu, J., Wu, Y., Chilukuri, R. V., Huang, L., Yamamoto, K., . . . Mita, K. (2017). Genomic adaptation to polyphagy and insecticides in a major East Asian noctuid pest. *Nature Ecology & Evolution, 1* (11), 1747-1756. doi:10.1038/s41559-017-0314-4

Chiba, T., & Kobayashi, M. (1985). Seasonal prevalence of the peach fruit moth, *Carposina niponensis* Walsingham, in the apple orchards in Iwate Prefecture. *Bulletin of the Iwate Horticultural Experiment Station, 6* , 1-14.

Cong, Q., Borek, D., Otwinowski, Z., & Grishin, N. V. (2015). Tiger swallowtail genome reveals mechanisms for speciation and caterpillar chemical defense. *Cell Rep, 10* (6), 910-919. doi:10.1016/j.celrep.2015.01.026

Dasmahapatra, K. K., Walters, J. R., Briscoe, A. D., Davey, J. W., Whibley, A., Nadeau, N. J., . . . Consortium, H. G. (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature, 487* (7405), 94-98. doi:10.1038/nature11041

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

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., & Aiden, E. L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems, 3* (1), 95-98. doi:10.1016/j.cels.2016.07.002

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., . . . Smart, A. (2018). The Pfam protein families database in 2019. *Nucleic Acids Research, 47* (D1), D427-D432.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., . . . Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research, 47* (D1), D427-D432. doi:10.1093/nar/gky995

Emms, D. M., & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology, 16* (157), 1-14. doi:10.1186/s13059-015-0721-2

Espinosa-Chaurand, D., Vega-Villasante, F., Carrillo-Farnes, O., & Nolasco-Soria, H. (2017). Effect of circadian rhythm, photoperiod, and molt cycle on digestive enzymatic activity of *Macrobrachium tenellum* juveniles. *Aquaculture, 479* , 225-232. doi:10.1016/j.aquaculture.2017.05.029

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research, 39* (suppl_2), W29-W37.

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 molesta*) at low temperature.*International Journal of Agriculture and Biology, 15* (1), 101-106.

Hu, B., Zhang, S. H., Ren, M. M., Tian, X. R., Wei, Q., Mburu, D. K., & Su, J. Y. (2019). The expression of Spodoptera exigua P450 and UGT genes: tissue specificity and response to insecticides. *Insect Science, 26* (2), 199-216. doi:10.1111/1744-7917.12538

Hua, B.-Z., Zeng, X.-H., & Zhang, H. (1998). Diapause of*Carposina sasakii* Matsumura (Lepidoptera Carposinidae) on various host plants. *Acta Universitatis Agriculturae Boreali-occidentalis, 26* (5), 25-29.

Hua, L., & Hua, B. Z. (1995). Preliminary study on the host-biotypes of peach fruit borer. *Acta Phytophylacica Sinica, 22* (2), 165-170.

Huang, K. X., Wang, Y. Z., Ye, Z. X., Zhang, N. X., & Zhang, L. Y. (1976). Influence of photoperiod and temperature on diapause of the peach fruit moth *Carposina sasakii* Matsumura. *Acta Entomologica Sinica, 19* (2), 149-156.

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., & Bork, P. (2017). Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper.*Molecular Biology and Evolution, 34* (8), 2115-2122. doi:10.1093/molbev/msx148 %J Molecular Biology and Evolution

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernandez-Plaza, A., Forslund, S. K., Cook, H., . . . Bork, P. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research, 47* (D1), D309-D314. doi:10.1093/nar/gky1085

Isobe, M., Kai, H., Kurahashi, T., Suwan, S., Pitchayawasin-Thapphasaraphong, S., Franz, T., . . . Nishida, H. (2006). The molecular mechanism of the termination of insect diapause, part 1: A timer protein, TIME-

EA4, in the diapause eggs of the silkworm*Bombyx mori* is a metallo-glycoprotein. *Chembiochem, 7* (10), 1590-1598. doi:10.1002/cbic.200600138

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

Kanost, M. R., Arrese, E. L., Cao, X., Chen, Y. R., Chellapilla, S., Goldsmith, M. R., . . . Blissard, G. W. (2016). Multifaceted biological insights from a draft genome sequence of the tobacco hornworm moth, Manduca sexta. *Insect Biochemistry and Molecular Biology, 76* , 118-147. doi:10.1016/j.ibmb.2016.07.005

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

Kim, D.-S., Lee, J.-H., & Yiem, M.-S. (2000). Spring Emergence Pattern ofCarposina sasakii(Lepidoptera: Carposinidae) in Apple Orchards in Korea and its Forecasting Models Based on Degree-Days.*Environmental Entomology, 29* (6), 1188-1198. doi:10.1603/0046-225x-29.6.1188

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology, 37* (8), 907-915. doi:10.1038/s41587-019-0201-4

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology, 14* (4), R36. doi:10.1186/gb-2013-14-4-r36

Kim, D. S., Lee, J. H., & Yiem, M. S. (2001). Temperature-dependent development of *Carposina sasakii* (Lepidoptera : Carposinidae) and its stage emergence models. *Environmental Entomology, 30* (2), 298-305.

Kim, Y. I., Kim, H. J., Kwon, Y. M., Kang, Y. J., Lee, I. H., Jin, B. R., . . . Seo, S. J. (2010). Modulation of MnSOD protein in response to different experimental stimulation in *Hyphantria cunea* .

Comparative Biochemistry and Physiology B-Biochemistry & Molecular Biology

157 (4), 343-350. doi:10.1016/j.cbpb.2010.08.003

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 adaptivek-mer weighting and repeat separation. *Genome Research, 27* (5), 722-736. doi:10.1101/gr.215087.116

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

Kozak, G. M., Wadsworth, C. B., Kahne, S. C., Bogdanowicz, S. M., Harrison, R. G., Coates, B. S., & Dopman, E. B. (2019). Genomic Basis of Circannual Rhythm in the European Corn Borer Moth. *Curr Biol, 29* (20), 3501-3509 e3505. doi:10.1016/j.cub.2019.08.053

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets.*Molecular Biology and Evolution, 33* (7), 1870-1874. doi:10.1093/molbev/msw054

Kwon, D., Kwon, H., Kim, D., & Yang, C. (2018). Larval species composition and genetic structures of Carposina sasakii, Grapholita dimorpha, and Grapholita molesta from Korea. *Bulletin of entomological research, 108* (2), 241-252.

Kwon, D. H., Kim, D. H., Kim, H. H., Lee, S. H., & Yang, C. Y. (2017). Genetic diversity and structure in apple-infesting pests of*Carposina sasakii* , *Grapholita dimorpha* and*Grapholita molesta* in Korea. *Journal of Asia-Pacific Entomology, 20* (1), 13-16. doi:10.1016/j.aspen.2016.11.002

Lagesen, K., Hallin, P., Rodland, E. A., Staerfeldt, H. H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research, 35* (9), 3100-3108. doi:10.1093/nar/gkm160

Lange, A., Beier, S., Huson, D. H., Parusel, R., Iglauer, F., & Frick, J. S. (2018). Genome Sequence of *Galleria mellonella* (Greater Wax Moth). *Genome Announcement, 6* (2), e01220-01217. doi:10.1128/genomeA.01220-17

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods, 9* (4), 357-359. doi:10.1038/nmeth.1923

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

Lu, S., Wang, J., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., . . . Marchler-Bauer, A. (2020). CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Research, 48* (D1), D265-D268. doi:10.1093/nar/gkz991

Lu, S., Yang, J., Dai, X., Xie, F., He, J., Dong, Z., . . . Li, X. (2019). Chromosomal-level reference genome of Chinese peacock butterfly (*Papilio bianor*) based on third-generation DNA sequencing and Hi-C analysis. *Gigascience, 8* (11). doi:10.1093/gigascience/giz128

Ma, W., Zhao, X., Yin, C., Jiang, F., Du, X., Chen, T., . . . Lin, Y. (2020). A chromosome-level genome assembly reveals the genetic basis of cold tolerance in a notorious rice insect pest, *Chilo suppressalis* . *Molecular Ecology Resources, 20* (1), 268-282. doi:10.1111/1755-0998.13078

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

Mitter, C., Davis, D. R., & Cummings, M. P. (2017). Phylogeny and evolution of Lepidoptera. *Annu Rev Entomol, 62* , 265-283. doi:10.1146/annurev-ento-031616-035125

Nishikawa, H., Iijima, T., Kajitani, R., Yamaguchi, J., Ando, T., Suzuki, Y., . . . Fujiwara, H. (2015). A genetic mechanism for female-limited Batesian mimicry in *Papilio* butterfly.*Nature Genetics, 47* (4), 405-409. doi:10.1038/ng.3241

Pearce, S. L., Clarke, D. F., East, P. D., Elfekih, S., Gordon, K. H. J., Jermiin, L. S., . . . Wu, Y. D. (2017). Genomic innovations, transcriptional plasticity and gene loss underlying the evolution and divergence of two highly polyphagous and invasive *Helicoverpa*pest species. *BMC Biology, 15* (1), 63. doi:10.1186/s12915-017-0402-6

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

Price, M. N., Dehal, P. S., & Arkin, A. P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution, 26* (7), 1641-1650. doi:10.1093/molbev/msp077

Ragland, G. J., Armbruster, P. A., & Meuti, M. E. (2019). Evolutionary and functional genetics of insect diapause: a call for greater integration. *Current Opinion in Insect Science*

36 , 74-81. doi:10.1016/j.cois.2019.08.003

Ramzanzadeh, F., Yeganeh, S., JaniKhalili, K., & Babaei, S. S. (2016). Effects of different photoperiods on digestive enzyme activities in rainbow trout (*Oncorhynchus mykiss*) alevin and fry.*Canadian Journal of Zoology, 94* (6), 435-442. doi:10.1139/cjz-2015-0180

Rane, R. V., Ghodke, A. B., Hoffmann, A. A., Edwards, O. R., Walsh, T. K., & Oakeshott, J. G. (2019). Detoxifying enzyme complements and host use phenotypes in 160 insect species. *Current Opinion in Insect Science, 31* , 131-138.

Roach, M. J., Schmidt, S. A., & Borneman, A. R. (2018). Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics, 19* (1), 460. doi:10.1186/s12859-018-2485-7

Sato, N., & Ishitani, M. (1976). Life-cycle of the peach fruit moth,*Carposina niponensis* Walsingham. *Bulletin of the Aomori Field Crops and Horticultural Experiment Station, 1* , 1-16.

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 Biology, 16* , 259. doi:10.1186/s13059-015-0831-x

Shan, X. J., Xiao, Z. Z., Huang, W., & Dou, S. Z. (2008). Effects of photoperiod on growth, mortality and digestive enzymes in miiuy croaker larvae and juveniles. *Aquaculture, 281* (1-4), 70-76. doi:10.1016/j.aquaculture.2008.05.034

Sim, C., & Denlinger, D. L. (2011). Catalase and superoxide dismutase-2 enhance survival and protect ovaries during overwintering diapause in the mosquito *Culex pipiens* . *Journal of Insect Physiology, 57* (5), 628-634. doi:10.1016/j.jinsphys.2011.01.012

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

Song, S. V., Downes, S., Parker, T., Oakeshott, J. G., & Robin, C. (2015). High nucleotide diversity and limited linkage disequilibrium in Helicoverpa armigera facilitates the detection of a selective sweep.*Heredity (Edinb), 115* (5), 460-470. doi:10.1038/hdy.2015.53

Stanke, M., & Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics, 19 Suppl 2* , ii215-225. doi:10.1093/bioinformatics/btg1080

Subala, S. P., & Shivakumar, M. S. (2017). Circadian variation affects the biology and digestive profiles of a nocturnal insect*Spodoptera litura* (Insecta: Lepidoptera). *Biological Rhythm Research, 48* (2), 207-226. doi:10.1080/09291016.2016.1251928

Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics, 25* (1), unit 4.10.

Toshima, A., Honma, K., Masaki, S. J. J. J. o. A. E., & Zoology. (1961). Factors Influencing the Seasonal Incidence and Breaking of Diapause in *Carposina niponensis* WALSHINGHAM. *Japanese Journal of Applied Entomology and Zoology, 5* (4), 260-269.

Vizueta, J., Sanchez-Gracia, A., & Rozas, J. (2019). BITACORA: A comprehensive tool for the identification and annotation of gene families in genome assemblies. *bioRxiv* , 593889. doi:10.1101/593889

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics, 33* (14). doi:10.1093/bioinformatics/btx153

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), e112963. doi:10.1371/journal.pone.0112963

Wan, F. H., Yin, C. L., Tang, R., Chen, M. H., Wu, Q., Huang, C., . . . Li, F. (2019). A chromosome-level genome assembly of *Cydia pomonella* provides insights into chemical ecology and insecticide resistance. *Nature Communications, 10* , doi.org/10.1038/s41467-41019-12175-41469. doi:ARTN 4237

10.1038/s41467-019-12175-9

Wang, J., Yu, Y., Li, L. L., Guo, D., Tao, Y. L., & Chu, D. (2015).*Carposina sasakii* (Lepidoptera: Carposinidae) in its native range consists of two sympatric cryptic lineages as revealed by mitochondrial COI gene sequences. *Journal of Insect Science, 15* (1), 1-6. doi:10.1093/jisesa/iev063

Wang, Y. Z., Li, B. Y., Hoffmann, A. A., Cao, L. J., Gong, Y. J., Song, W., . . . Wei, S. J. (2017). Patterns of genetic variation among geographic and host-plant associated populations of the peach fruit moth*Carposina sasakii* (Lepidoptera: Carposinidae). *BMC Evolutionary Biology, 17* (1), 265. doi:10.1186/s12862-017-1116-7

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

Xiang, H., Liu, X., Li, M., Zhu, Y. n., Wang, L., Cui, Y., . . . Zhan, S. (2018). The evolutionary road from wild moth to domestic silkworm.*Nature Ecology & Evolution, 2* (8), 1268-1279. doi:10.1038/s41559-018-0593-4

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

Xu, Q. G., & Hua, B. Z. (2004). RAPD analysis on the speciation in host races of *Carposina sasakii* Matsumura (Lepidoptera: Carposinidae).*Acta Entomologica Sinica, 47* (3), 379-383.

Yang, P., Tanaka, H., Kuwano, E., & Suzuki, K. (2008). A novel cytochrome P450 gene (CYP4G25) of the silkmoth *Antheraea yamamai* : Cloning and expression pattern in pharate first instar larvae in relation to diapause. *Journal of Insect Physiology, 54* (3), 636-643. doi:10.1016/j.jinsphys.2008.01.001

You, M. S., Yue, Z., He, W. Y., Yang, X. H., Yang, G., Xie, M., . . . Wang, J. (2013). A heterozygous moth genome provides insights into herbivory and detoxification. *Nature Genetics, 45* (2), 220-225.

Zhan, S., Merlin, C., Boore, J. L., & Reppert, S. M. (2011). The monarch butterfly genome yields insights into long-distance migration.*Cell, 147* (5), 1171-1185. doi:10.1016/j.cell.2011.09.052

Zhang, B., Peng, Y., Zhao, X. J., Hoffmann, A. A., Li, R., & Ma, C. S. (2016). Emergence of the overwintering generation of peach fruit moth (*Carposina sasakii*) depends on diapause and spring soil temperatures. *Journal of Insect Physiology, 86* , 32-39. doi:10.1016/j.jinsphys.2015.12.007

Zhang, S., Shen, S., Peng, J., Zhou, X., Kong, X., Ren, P., . . . Zhang, Z. (2020). Chromosome-level genome assembly of an important pine defoliator, *Dendrolimus punctatus* (Lepidoptera; Lasiocampidae).*Molecular Ecology Resources, 20* (4), 1023-1037.

Zhao, L., & Shi, L. (2009). Metabolism of hydrogen peroxide in univoltine and polyvoltine strains of silkworm (*Bombyx mori*).*Comparative Biochemistry and Physiology B-Biochemistry & Molecular Biology, 152* (4), 339-345. doi:10.1016/j.cbpb.2008.12.014

Tables

Table 1 Features of chromosome-level genomes in the Lepidoptera

Features	Csas	Cpom	Tni	Bmor	Slit	Mcin	Hmel
Genome size (Mb)	401.67	772.89	368.2	431.7	438.32	393	269
Karyotype	2n=64	2n=56	2n=54	2n=56	2n=62	2n=62	2n=42
No. contigs	275	2221	26,605	15,018	13,636	49,851	NA
No. scaffolds	NA	1717	6181	7397	3597	8262	3807
No. CHR*	31A+Z+W	27A+Z+W	26A+Z+W	27A+Z	30A+Z	30A+Z	20A+Z
Contig N50 (kb)	2620	862.49	621.9	15.5	68.35	13	51
Scaffold N50 (Mb)	NA	8.92	14.2	3.7	0.915	0.119	0.277
BUSCO genes (%)	98.20%	98.5	97.8	97.7	98.3	91.5	97.4

Features	**Csas**	**Cpom**	**Tni**	**Bmor**	**Slit**	**Mcin**	**Hmel**
Repeat (%)	11.33	42.87	20.5	43.6	31.83	28	24.94
G+C (%)	36.96	37.43	35.6	37.3	36.5	33.0	NA
No. genes	23,227	17,184	14,043	14,623	15,317	16,667	12,669

Csas, *Carposina sasakii* ; Cpom, *Cydiapomonella* ; Tni, *Trichoplusia ni* ; Bmor,*Bombyx mori* ; Slit, *Spodoptera litura* ; Mcin,*Melitaea cinxia* ; Hmel, *Heliconius melpomene* ; * A represents auto chromosome; Z and W represent sex chromosomes; NA, not available. Data for all species except for Csas were summarized by Wan et al. (2019).

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

Species	P450	GST	CCE	ABC	UGT	Reference for genome
Carposina sasakii	96	77	63	104	28	This study
Plutella xylostella	74*	11*	55*	97*	38	You et al. (2013)
Cydia pomonella	137**	30**	73**	47**	30**	Wan et al. (2019)
Danaus plexippus	107	35	73	76	47	Zhan et al. (2011)
Trichoplusia ni	143	51	122	71	68	W. Chen et al. (2019)
Spodoptera litura	138*	47*	110*	54*	64	Cheng et al. (2017)
Helicoverpa armigera	122	57	105	76	54	Song, Downes, Parker, Oakeshott, and Robin (2015)
Bombyx mandarina	94	37	94	64	48	Xiang et al. (2018)
Bombyx mori	83*	26*	76*	51*	50	Xia et al. (2004)
Manduca sexta	97*	17*	86*	47*	51	Kanost et al. (2016)
Galleria mellonella	137	44	75	72	58	Lange et al. (2018)
Ostrinia furnacalis	126	48	115	112	46	Ma et al. (2020)

*, data from S. Zhang et al. (2020); **, data from Wan et al. (2019). The other data were manually identified in our study.

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).

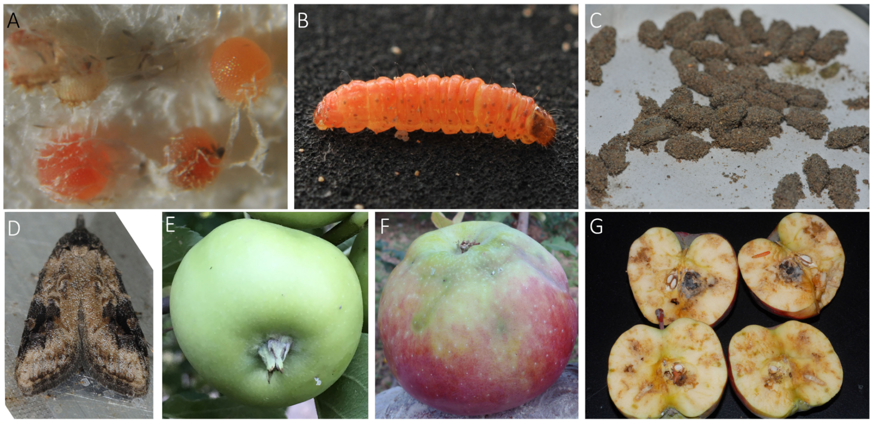
Fig. 2 Genome features of *Carposina sasakii.* (a) GenomeScope analysis of genome size, heterozygosity and duplicate rate using k-mers (K = 17) count histogram, indicating a genome size of 338.52 Mb, a heterozygosity of 1.06%, and a duplication rate of 2.06%; (b) Genome-wide all-by-all Hi-C interaction identified 32 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 Comparative genomics of *Carposina sasakii.* (a) 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. (b) Orthogroups shared by five Lepidoptera species of *Carposina sasakii* ,*Cydia pomonella* , *Bombyx mori* , *Ostrinia furnacalis* and *Helicoverpa armigera.*

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

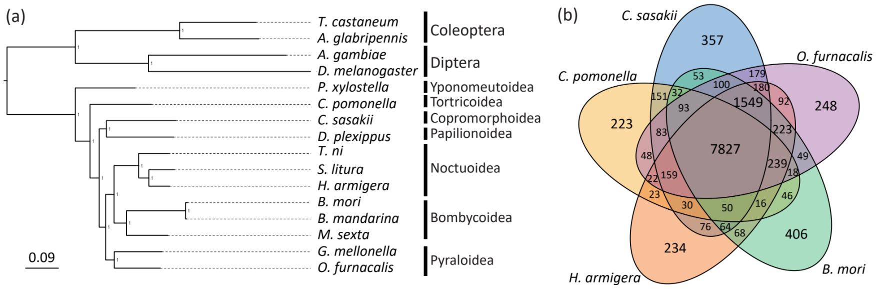
domains on protein sequences. Species and their taxonomic status are shown on the left: Tcas, *Tribolium castaneum* ; Agla,*Anoplophora glabripennis* ; Agam, *Anopheles gambiae* ; Dmel,*Drosophila melanogaster* ; Pxy, *Plutella xylostella* ; Cpom,*Cydia pomonella* ; Csas, *Carposina sasakii* ; Dple,*Danaus 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* .

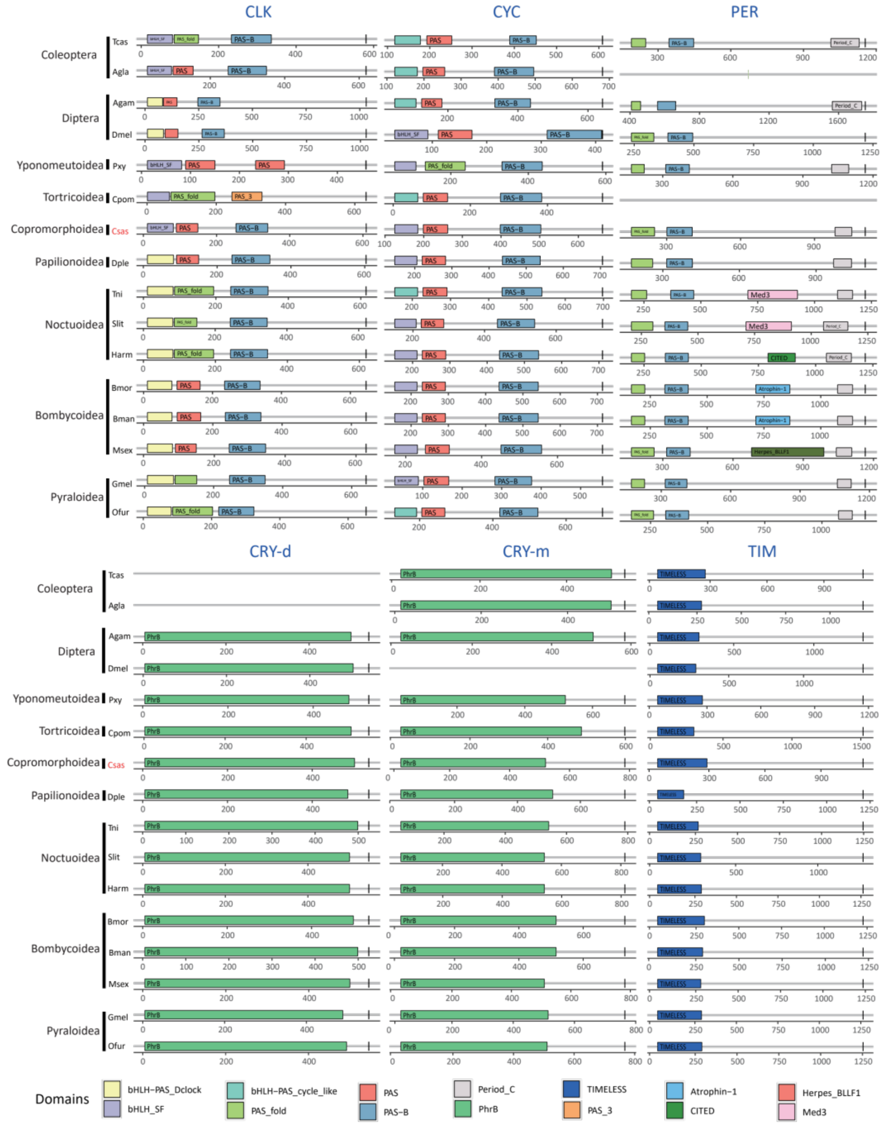
Fig. 5 Phylogenetic relationships of two PAS domains in three circadian genes: *period* (per), *Clock* (clk) and*cycle* (cyc). Each tip is labeled by the name of domain, gene and species. Abbreviations of species are same as in Fig. 4. Six clades shaded in different color reveal two domains of three genes, while one domain of *clk* gene has two different type among species. Tips in red show the position of *Carposina sasakii* .

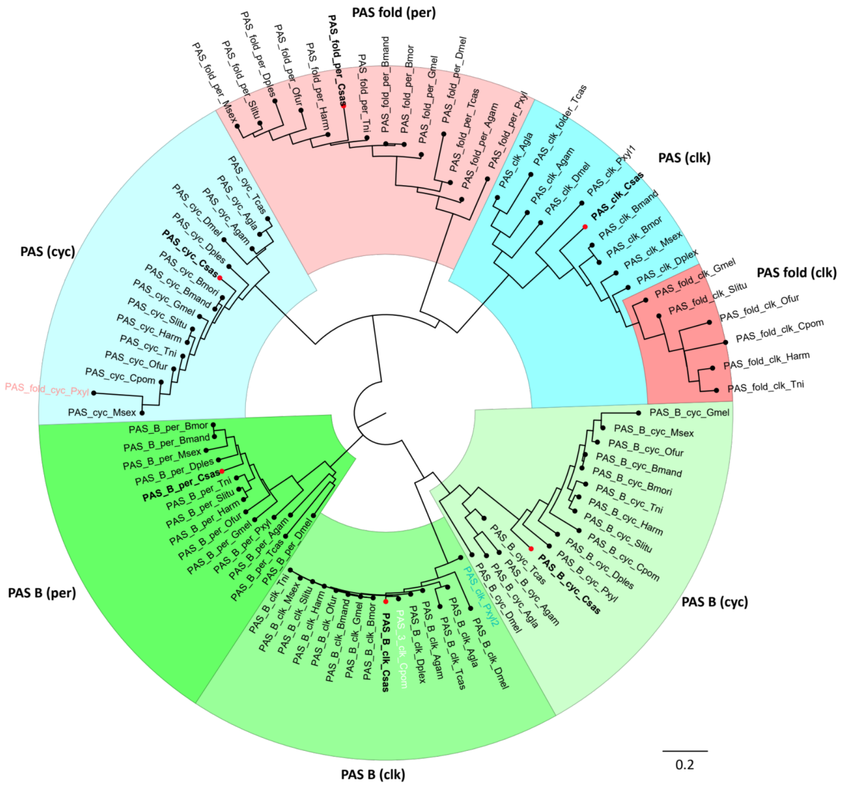


Hosted file

`Fig.2_1.pdf` available at https://authorea.com/users/350968/articles/475653-chromosome-level-genome-of-the-peach-fruit-moth-carposina-sasakii-lepidoptera-carposinidae-provides-a-resource-for-evolutionary-studies-on-moths




CLK
CYC
PER
CRY-d
CRY-m
TIM
Coleoptera
Tcas
Agla
Diptera
Agam
Dmel
Yponomeutoidea
Pxy
Tortricoidea
Cpom
Copromorphoidea
Csas
Papilionoidea
Dple
Noctuoidea
Tni
Slit
Harm
Bombycoidea
Bmor
Bman
Msex
Pyraloidea
Gmel
Ofur
bHLH_SF
PAS_fold
PAS
PAS-B
PAS_3
Period_C
Med3
CITED
Atrophin-1
Herpes_BLLF1
PhrB
TIMELESS
Domains
🟨 bHLH-PAS_Dclock
🟪 bHLH_SF
🟦 bHLH-PAS_cycle_like
🟩 PAS_fold
🟥 PAS
🟦 PAS-B
⬜ Period_C
🟩 PhrB
🟦 TIMELESS
🟧 PAS_3
🟦 Atrophin-1
🟩 CITED
🟥 Herpes_BLLF1
🟪 Med3


PAS fold (per)
PAS_fold_per_Msex
PAS_fold_per_Slitu
PAS_fold_per_Dples
PAS_fold_per_Ofur
PAS_fold_per_Harm
PAS_fold_per_Csas
PAS_fold_per_Tni
PAS_fold_per_Bmand
PAS_fold_per_Bmor
PAS_fold_per_Gmel
PAS_fold_per_Dmel
PAS_fold_per_Tcas
PAS_fold_per_Agam
PAS_fold_per_Pxyl
PAS (clk)
PAS_clk_Agla
PAS_clk_folder_Tcas
PAS_clk_Agam
PAS_clk_Dmel
PAS_clk_Pxyl1
PAS_clk_Csas
PAS_clk_Bmand
PAS_clk_Bmor
PAS_clk_Msex
PAS_clk_Dplex
PAS fold (clk)
PAS_fold_clk_Gmel
PAS_fold_clk_Slitu
PAS_fold_clk_Ofur
PAS_fold_clk_Cpom
PAS_fold_clk_Harm
PAS_fold_clk_Tni
PAS_B_cyc_Gmel
PAS_B_cyc_Msex
PAS_B_cyc_Ofur
PAS_B_cyc_Bmand
PAS_B_cyc_Bmori
PAS_B_cyc_Tni
PAS_B_cyc_Harm
PAS_B_cyc_Slitu
PAS_B_cyc_Cpom
PAS_B_cyc_Dples
PAS_B_cyc_Pxyl
PAS_B_cyc_Csas
PAS_B_cyc_Tcas
PAS_B_cyc_Agam
PAS_B_cyc_Agla
PAS_B_cyc_Dmel
PAS_clk_Pxyl2
PAS B (cyc)
PAS_B_clk_Dmel
PAS_B_clk_Agla
PAS_B_clk_Tcas
PAS_B_clk_Agam
PAS_B_clk_Dplex
PAS_3_clk_Cpom
PAS_B_clk_Csas
PAS_B_clk_Bmor
PAS_B_clk_Gmel
PAS_B_clk_Bmand
PAS_B_clk_Ofur
PAS_B_clk_Harm
PAS_B_clk_Slitu
PAS_B_clk_Msex
PAS_B_clk_Tni
PAS B (clk)
PAS_B_per_Dmel
PAS_B_per_Tcas
PAS_B_per_Agam
PAS_B_per_Pxyl
PAS_B_per_Gmel
PAS_B_per_Ofur
PAS_B_per_Harm
PAS_B_per_Slitu
PAS_B_per_Tni
PAS_B_per_Csas
PAS_B_per_Dples
PAS_B_per_Msex
PAS_B_per_Bmand
PAS_B_per_Bmor
PAS B (per)
PAS_cyc_Msex
PAS_fold_cyc_Pxyl
PAS_cyc_Cpom
PAS_cyc_Ofur
PAS_cyc_Tni
PAS_cyc_Harm
PAS_cyc_Slitu
PAS_cyc_Gmel
PAS_cyc_Bmand
PAS_cyc_Bmori
PAS_cyc_Csas
PAS_cyc_Dples
PAS_cyc_Dmel
PAS_cyc_Agam
PAS_cyc_Agla
PAS_cyc_Tcas
PAS (cyc)
0.2