


RESEARCH ARTICLE

Open Access



Genome of the webworm *Hyphantria cunea* unveils genetic adaptations supporting its rapid invasion and spread

Qi Chen^{1,2†}, Hanbo Zhao^{1,2†}, Ming Wen^{1,2}, Jiabin Li^{1,2}, Haifeng Zhou^{1,2}, Jiatong Wang^{1,2}, Yuxin Zhou^{1,2}, Yulin Liu^{1,2}, Lixin Du^{1,2}, Hui Kang^{1,2}, Jian Zhang³, Rui Cao⁴, Xiaoming Xu⁵, Jing-Jiang Zhou^{1,2,6}, Bingzhong Ren^{1,2} and Yinliang Wang^{1,2*} 

Abstract

Background: The fall webworm *Hyphantria cunea* is an invasive and polyphagous defoliator pest that feeds on nearly any type of deciduous tree worldwide. The silk web of *H. cunea* aids its aggregating behavior, provides thermal regulation and is regarded as one of causes for its rapid spread. In addition, both chemosensory and detoxification genes are vital for host adaptation in insects.

Results: Here, a high-quality genome of *H. cunea* was obtained. Silk-web-related genes were identified from the genome, and successful silencing of the silk protein gene *HcunFib-H* resulted in a significant decrease in silk web shelter production. The CAFE analysis showed that some chemosensory and detoxification gene families, such as *CSPs*, *CCEs*, *GSTs* and *UGTs*, were expanded. A transcriptome analysis using the newly sequenced *H. cunea* genome showed that most chemosensory genes were specifically expressed in the antennae, while most detoxification genes were highly expressed during the feeding peak. Moreover, we found that many nutrient-related genes and one detoxification gene, *HcunP450* (CYP306A1), were under significant positive selection, suggesting a crucial role of these genes in host adaptation in *H. cunea*. At the metagenomic level, several microbial communities in *H. cunea* gut and their metabolic pathways might be beneficial to *H. cunea* for nutrient metabolism and detoxification, and might also contribute to its host adaptation.

Conclusions: These findings explain the host and environmental adaptations of *H. cunea* at the genetic level and provide partial evidence for the cause of its rapid invasion and potential gene targets for innovative pest management strategies.

Keywords: Genome, Metagenome, Genetics, Fall webworm, Molecular evolution, Adaptation, Gene expansion

* Correspondence: wangyl392@nenu.edu.cn

†Qi Chen and Hanbo Zhao contributed equally to this work.

¹Jilin Provincial Key Laboratory of Animal Resource Conservation and Utilization, Northeast Normal University, Changchun, Jilin, China

²Key Laboratory of Vegetation Ecology, MOE, Northeast Normal University, Changchun, China

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

The fall webworm, *Hyphantria cunea* Drury (Erebidae: Hyphantria), is a polyphagous pest species in forest and agricultural ecosystems; where its larvae feed on most deciduous tree leaves [1]. When trees are infested, the fall webworm consumes nearly all leaves and causes great ecological and economic impact to the forest industry [2]. *H. cunea* is also an invasive pest, native to North America, but has spread globally in the past seven decades [3]. Behavioural, physiological and ecological adaptations present in this species are believed to contribute to its rapid spread.

First, the fall webworm has an extremely wide range of host plants and been reported to forage on more than 600 plant species, covering nearly all types of deciduous trees, especially mulberry, boxelder, walnut, sycamore, apple, plum, cherry, and elm [4]. Insect host selection is regulated by the chemosensory systems [5], especially for polyphagous herbivores [6–8]. Insect chemosensory systems consist of several gene families, including odorant receptor (*OR*), gustatory receptor (*GR*), ionotropic receptor (*IR*), chemosensory protein (*CSP*) and odorant binding protein (*OBP*) families. These genes encode proteins that participate in host plant detection and sexual communication [9–12]. Previous investigations have suggested that the large expansions in chemosensory gene families are a possible adaptation mechanism which enables polyphagy in the lepidopteran insect *Spodoptera frugiperda* [13] and other taxa such as *Apis mellifera*, *Bombyx mori* and *Bemisia tabaci* [9, 14–18]. Thus, chemosensory genes were further examined in this study to explore the roles of these genes in host plant adaptation of *H. cunea*. In addition, several studies have shown that the host ranges of insects are determined by their detoxification abilities [19, 20], which also contribute to adaptation to polyphagy in insect herbivores [13, 21]. Therefore, detoxification genes such as UDP-glycosyltransferase (*UGT*), glutathione S-transferase (*GST*), carboxyl/choline esterases (*CCE*), ATP-binding cassette transporter (*ABC*) and cytochrome P450 (*P450*) were screened from the transcriptome and metagenome datasets of *H. cunea* and analyzed for differential expression and positive selection.

Second, the fall webworm has a high reproductive capacity and a strong tolerance of extreme environments, including a wide range of temperatures (–16 °C to 40 °C) and starvation (the larvae of fall webworm can live without food for more than 10 days) [22]. Numerous studies have found that the gut bacteria of insects play crucial roles in environmental adaptation by their insect hosts [23–25]. Gut microbes with a mutualistic relationship to their hosts contribute to preventing pathogen growth in insects [26]. For example, the gut bacteria of the desert locust *Schistocerca gregaria* could protect the

locust gut from colonization by an insect pathogenic bacterium, *Serratia marcescens* [27]. Furthermore, gut microbial partnerships could help their insect hosts proliferate under a range of temperatures [28], conferring cold tolerance [29] and heat stress tolerance [30, 31]. Meanwhile, some gut bacteria and the natural products extracted from bacteria are used for pest control [23, 32]. Therefore, to gain new insights into the environmental adaptations of the fall webworm at the microbiome level, the compositional diversity of the gut microbiota in *H. cunea* was also investigated by metagenomic analysis in this study.

Finally, *H. cunea* larvae aggregate by creating silk webs on tree branches, this social behavior provides temperature regulation and protects them from predators [33, 34]. In most Lepidopteran species, the silk is composed of two major silk proteins, fibroin and sericin [35–37]. The fibroins form filaments, and the sericins seal and glue the filaments into fibers [37]. In caddisflies, the phosphorylation of fibroins was found to contribute to larval adaptation to aquatic habitats, suggesting that fibroin might be involved in environmental adaptation among silk-spinning insects [38]. Thus, we annotated in the *H. cunea* genome and identified genes from the silk gland, especially the silk proteins (fibroins and sericins) to explore the functions of these genes in *H. cunea*.

With the explosive growth of bioinformatics and sequencing technologies, many insect genomes have been sequenced and provided comprehensive information on the phylogeny, evolution, population geography, gene function and genetic adaptation of these insects. In Lepidoptera, at least ten species' genomes have been sequenced and published [39–46]. Wu et al. had performed a genome study on *Hyphantria cunea* and provided some insights into the rapid adaptation of the fall webworm to changing environments and host plants [47], in this study, a higher quality genome sequence of *H. cunea* was obtained by using a mix of PacBio and Illumina platform. Moreover, some evidences suggested that the gut bacteria of insects played essential roles in the adaptation of insects to their host plant [23–25], thus a further metagenomic analysis was performed in *H. cunea*, the results might provided us a better understanding of its rapid spread and also some potential gene targets for developing new methods to manage this worldwide pest.

Results

Overview of genome assembly and annotation

The genome survey with k-mer analysis (Figure S1) showed that there is a small peak in depth = 22 which represented the heterozygous sequences, while the

average k-mer depth was 45, and the peak indepth = 90 indicates the repetitive sequences. As a results, the tentative genome size of *H. cunea* was 563.96 Mb with a low heterozygosity of 0.23% and repetitive elements of 36.20% of the whole genome (Figure S1 and Table S1).

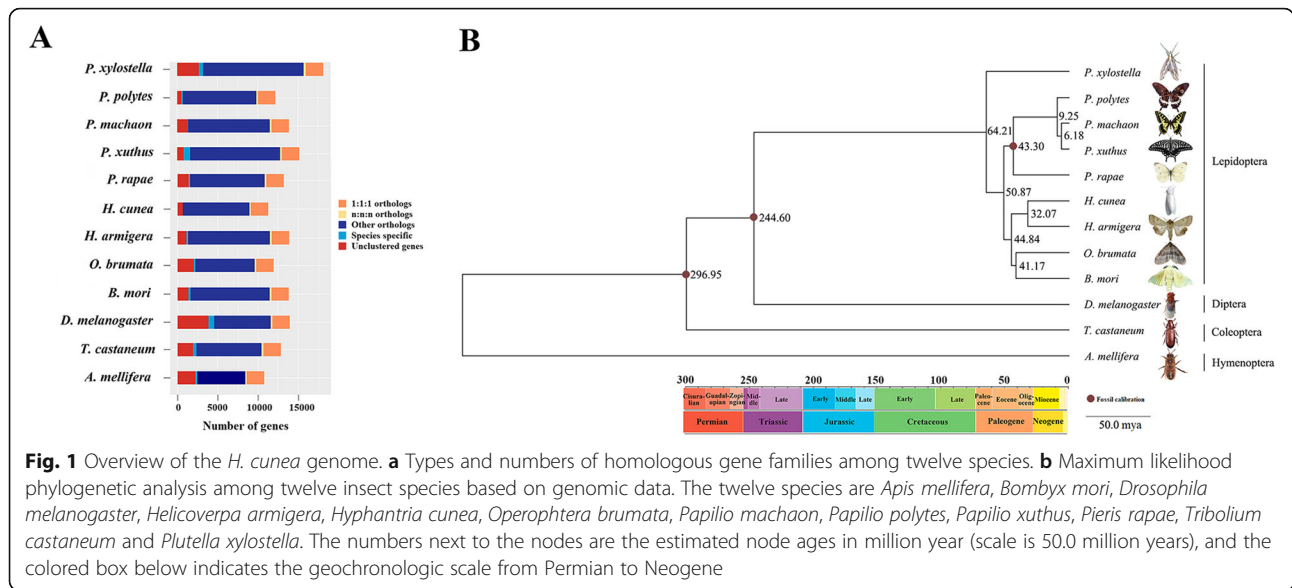
The generated genome assembly of *H. cunea* comprises a 559.30 Mb sequence with a 3.09 Mb contig N50. It contains 198.97 Mb of repetitive elements that occupy 35.71% of the genome. After correction with RNA sequencing data from 12 samples of different tissues and stages of *H. cunea*, we obtained 15,319 genes using three gene prediction strategies (Figure S2), 94.42% of which could be annotated and enriched by the GO and KOG databases (Figures S3 and S4), and the distribution of Nr homologous genes with the *H. cunea* genome in insect species was showed in Figure S5. Moreover, 637 tRNAs, 71 rRNAs, 48 miRNAs and 300 pseudogenes were predicted from the Rfam and miRBase databases by the Infernal, tRNAscan-SE and GenBlastA software (Table S2). Further analyses showed that 94.54 and 92.96% eukaryotic conserved genes were found in the genome of *H. cunea* by CEGMA and BUSCO, respectively, suggesting that the genome sequence we obtained was largely complete (Tables S3-S6). The genome of *H. cunea* possesses a comparatively longer contig N50 among all genomes of Lepidoptera species sequenced so far, the top 4 are as follows: *Operophtera brumata* (6.38 Mb) [40], *Spodoptera frugiperda* (5.6 Mb) [48], *Papilio bianor* (5.5 Mb) [49], and *H. cunea* (3.09 Mb), further confirming the high quality of the genome sequence of *H. cunea* (Table 1). Homology analysis of the *H. cunea* genome led to the identification of 2142 pairs of one-to-one single-copy orthologs among twelve species. This ortholog dataset was used for further studies described below. Only 27 genes were specific to *H. cunea*, which is the smallest species-specific number among the eight lepidopteran species (Fig. 1a).

Phylogeny of Lepidoptera

RAXML was used to construct a maximum likelihood phylogenetic tree using the 2142 single-copy orthologs among twelve insects whose genome sequences were available; eight Lepidoptera were included, while Hymenoptera (*A. mellifera*), Coleoptera (*T. castaneum*) and Diptera (*D. melanogaster*) were used as outgroups. The results showed that all nodes were supported by strong bootstrap values of 100%, and the topology of the higher taxa was consistent with those of previous phylogenetic studies [50, 51]. The results revealed that Lepidoptera was closer to Diptera, while Hymenoptera was located at the basal branch and formed a single clade (Fig. 1b). Within Lepidoptera, Papilionoidea (butterflies) formed a single clade, and *P. xylostella* (Yponomeutoidea) was separated from other moth taxa (Noctuoidea, Geometroidea and Bombycoidea). *H. cunea* was shown to be most closely related to *H. armigera*, which also belongs to the superfamily Noctuoidea. These results are in agreement with those obtained from the phylogenetic studies of Lepidoptera based on morphology and molecular data [52, 53]. The phylogenetic analysis indicated that Lepidoptera diverged from Diptera approximately 244.60 million years ago, which is consistent with the previously reported divergence time [50]. In Lepidoptera, the divergent time between the moths and butterflies in our study and was at Paleogene period, which is consist with Kawahara's work, moreover, the genetic relationship between GEOMETROIDEA and BOMBYCOIDEA were close related, and they were grouped together with NOCTUOIDEA as well [54]. *H. cunea* and *H. armigera* were estimated to have diverged at the Eocene-Oligocene boundary with a divergence time of approximately 32.07 million years ago. The period from the late Eocene to early Oligocene has been considered as an important transition time and a link between the archaic world of the tropical Eocene and the more modern ecosystems of the Miocene [55].

Table 1 Overview of sequenced lepidopteran genomes

Species	Assembly size (MB)	Protein-coding Gene number	Contig N50 (MB)	GC (%)	Intron (%)	Repeat (%)	Protein number	Pseudogenes
<i>Plutella xylostella</i>	393.47	19,340	1.85	39.8	30.70	34	21,661	145
<i>Papilio polytes</i>	227.02	13,301	4.78	34	24.8	n.a.	16,620	66
<i>Papilio machaon</i>	278.42	14,850	0.08	34.4	n.a.	n.a.	17,745	102
<i>Papilio xuthus</i>	243.23	15,322	0.49	34.9	45.5	n.a.	21,602	232
<i>Pieris rapae</i>	245.87	13,152	1.42	33	33.3	22.7	18,966	70
<i>Hyphantria cunea</i>	559.30	15,319	3.09	36.57	29.1	35.71	18,207	300
<i>Helicoverpa armigera</i>	337.07	15,081	2.35	37.5	39.3	14.6	21,035	65
<i>Operophtera brumata</i>	638.21	16,912	2.88	37.8	17.7	53.5	16,912	n.a.
<i>Bombyx mori</i>	481.82	16,166	1.55	38.8	16.3	43.6	22,571	64



Expansion of chemosensory and detoxification gene families

To further explore host adaptation, the *H. cunea* gene families related to chemosensory abilities (*ORs*, *GRs*, *IRs*, *CSPs* and *OBPs*) were studied. With the combination of de novo assembly, homology-based search and RNA sequencing annotation, 72 *ORs*, 46 *GRs*, 66 *OBPs*, 20 *CSPs* and 21 *IRs* were identified in the *H. cunea* genome (Table 2). This result increased the number of chemosensory genes in *H. cunea* from the previous identifications via antennal transcriptome studies, which reported 52 *ORs*, 9 *GRs*, 30 *OBPs*, 17 *CSPs* and 14 *IRs* [56]. For the gene families related to detoxification, 32 *UGTs*, 25 *GSTs*, 75 *CCEs*, 95 *ABCs* and 109 *P450s* were identified using the same strategy as above (Table 2). The numbers of chemosensory and detoxification genes in *H. cunea* were further compared with those of some lepidopteran insects (Table 2) [46].

Gene family expansion/contraction analyses showed that the *CSP*, *CCE*, *GST* and *UGT* gene families were expanded in *H. cunea* compared to the tested Lepidopteran species, as the divergence sizes were all significantly lower than the species sizes for these genes (Table 3). *CSPs* contribute to transportation, sensitivity and possibly the selectivity of the insect olfactory system [10]. In our study, an expansion of *CSPs* was detected, suggesting that they might relate to host plant selection of *H. cunea*, but much more testing is required. Among the detoxification gene families, *UGT*, *CCE* and *GST* families were found to be expanded in *H. cunea* (Table 3). Some studies also found that in some polyphagous species in Noctuoidea *GSTs* and *CCEs* were greatly expanded, such as *H. zea*, *H. armigera* and *S. litura* [45, 46].

Other major expanded gene families were hemolymph protein [57], cecropin A [58], serine protease [59], apolipoporphins [60], DNA helicase [61], insulin-like growth

Table 2 The number of chemosensory and detoxification genes of *H. cunea* and other insect

Gene name	<i>Hyphantria. cunea</i>	<i>Papilio machaon</i>	<i>Operophtera brumata</i>	<i>Helicoverpa armigera</i>	<i>Bombyx mori</i>
CSPs	20	23	13	22	21
OBPs	66	29	15	30	43
ORs	72	51	29	74	73
GRs	46	10	9	19	76
IRs	21	33	31	52	31
P450s	109	112	133	112	83
GSTs	25	11	11	11	26
CCEs	74	53	42	53	73
APNs	24	17	28	17	14
ABCs	78	120	90	124	51
UGTs	32	39	11	39	45

Table 3 Gene families expanded in *H. cunea* as calculated by CAFE

Divergence size	Species size	P-value	Gene ID	Annotation
2	4	0.006	EVM0007968 EVM0014260 EVM0001423.	Adenosine deaminase-related growth factor A
2	4	<1e-7	EVM0003438 EVM0001582 EVM0008984 EVM0011535	ATP-dependent helicase YHR031C
2	3	0.002	EVM0003192 EVM0011563 EVM0002688	Cytoskeleton
1	2	<1e-7	EVM0000776	Kazal-type serine proteinase inhibitor 1; gag-like protein
2	3	0.001	EVM0015065 EVM0001703 EVM0008164	Uncharacterized protein
1	2	<1e-7	EVM0000473 EVM0004248	Retroelement polyprotein
2	3	0.021	EVM0008576 EVM0007758 EVM0001530	Carboxyl/choline esterase CCE033a
1	2	0.047	EVM0012638 EVM0014428	PREDICTED: serine/threonine-protein kinase SMG1-like
1	2	0.041	EVM0011073 EVM0014934	Uncharacterized protein
1	2	<1e-7	EVM0014320 EVM0009912	Uncharacterized protein LOC103572275
1	2	0.021	EVM0001502 EVM0011366	Hypothetical protein KGM_05165
1	2	<1e-7	EVM0009792 EVM0013551	Uncharacterized protein LOC101742343
1	3	<1e-7	EVM0002515 EVM0008850	Uncharacterized protein
1	2	0.021	EVM0014968 EVM0000652	Lysosomal-trafficking regulator-like
1	2	0.001	EVM0013380 EVM0003562	Chemosensory protein precursor
1	2	0.024	EVM0008361 EVM0005241	Insulin-like growth factor 2 mRNA-binding protein
1	2	<1e-7	EVM0014740 EVM0000171	GTPase-activating protein pac-1-like
1	2	<1e-7	EVM0012402 EVM0000677	Hemolymph protein 14
1	2	0.001	EVM0013787 EVM0001757	Glutathione S-transferase
1	2	0.005	EVM0003715 EVM0006371	Hypothetical protein 3 - cabbage looper transposon TED
1	2	0.001	EVM0009126 EVM0008595	Retinol dehydrogenase 11-like
1	2	0.037	EVM0011934 EVM0013760	Apolipoporphins
1	2	0.027	EVM0012995 EVM0001222	S-antigen protein
1	2	<1e-7	EVM0007537 EVM0012615	Cecropin A
1	2	<1e-7	EVM0004477 EVM0000882.	UDP-glycosyltransferase
1	2	<1e-7	EVM0007937 EVM0005030	Endonuclease and reverse transcriptase-like protein
1	2	0.023	EVM0014113 EVM0014355	Amino acid transport and metabolism; Serine protease 24
1	3	0.012	EVM0012395 EVM0001096	ATP-dependent DNA helicase MER
1	2	0.044	EVM0003209 EVM0014358	Proline-rich protein; lebecin-like protein
1	3	0.001	EVM0010172 EVM0008560 EVM0012239	Uncharacterized protein
1	2	<1e-7	EVM0001934 EVM0003728	Hypothetical protein 2 - cabbage looper transposon TED
1	2	<1e-7	EVM0009953 EVM0015050	Piggybac transposable element-derived
1	2	<1e-7	EVM0010118 EVM0011622	Receptor guanylate cyclase
1	2	0.014	EVM0006212 EVM0007033	Yolk protein 2
1	2	0.021	EVM0001362 EVM0000363	Uncharacterized protein
1	2	0.021	EVM0011788 EVM0007752	Pickpocket protein 28-like
1	3	<1e-7	EVM0000920 EVM0007834 EVM0004965	Uncharacterized protein

factor [62], and yolk proteins [63, 64] (Table 3). These gene families are supported to be involved in immunity, growth and development, biomacromolecule metabolism and reproduction in insects [57–59, 65–68].

DEG analysis in different stages and tissues

To further study the chemosensory and detoxification gene families that were found to be expanded, transcriptome studies on these genes were performed to explore

their expression profiles in different developmental stages and tissues. The analysis of differential gene expression by pairwise comparison led to the identification of 8232 DEGs within the different stages RNA (eggs, second instar larvae, fourth instar larvae, pupae, and male and female adults), and 7733 DEGs within the different tissues RNA (head, thorax, leg, abdomen, antenna, and female sexual glands). Then these two DEG datasets were combined, and the duplicated sequences were removed to create a final dataset of 10,348 DEGs (Table S7). The relative expression levels of these DEGs in different tissues and stages as indicated by \log_{10} FPKM values were shown in as the Box plot in Figure S6, and the numbers of alternative splicing events was showed in Figure S7. The expression of DEG gene families (*CSPs*, *GSTs*, *CCEs* and *UGTs*) was transformed into an expression heatmap and presented in Fig. 2 to better compare their expression levels in different tissues and developmental stages. Nine of the 20 *CSPs* were grouped together and specifically expressed in the antennae, four *CSPs* were highly expressed in pupae relative to other stages, while two *CSPs* were specifically expressed in the sex gland (Fig. 2a). In the expanded detoxification gene families *CCE*, *GST* and *UGT*, some genes were highly expressed in the fourth larval instar (Fig. 2b, c and d), which is the peak period of *H. cunea* foraging behavior [1].

Positive selection on genes related to nutrient metabolism and detoxification

Next, a positive selection analysis based on the homolog genes was performed on the genome of *H. cunea* to gain a better understanding of the mechanisms in its host selection. The branch-site model showed that 39 genes were under significant positive selection pressure (LRT, $p < 0.05$), of which 13 were nutrient regulation genes reported to be involved in the metabolism of lipids, carbohydrates, vitamins and amino acids (Table 4). Many studies have shown that nutrient regulation in herbivorous insects is shaped by natural selection [69, 70]. Significant positive selection pressure was also detected in *HcunP450* (EVM0009687), a member of the major detoxification-related gene family *P450* (Table 4), consistent with a previous study reported that *P450s* could mediate insect resistance to many classes of insecticides [71]. *HcunP450* was most similar to the cytochrome *P450 CYP306A1* of the cotton bollworm *H. armigera* (AID54855.1), with 81.63% identity at the amino acid level. The expression of *HarmP450 CYP306A1* was found to be induced by 2-tridecanone, and to mediate cotton bollworm development [72]. The *CYP306A1* gene family was also shown to play an essential role in ecdysteroid biosynthesis during insect development [73], in

fluoride resistance of *B. mori* [74]. Thus, the positive selection of *HcunP450 CYP306A1* might reflect the rapid development of insecticide resistance in *H. cunea*. However, it is needed to determine whether it is caused by long-term host adaptation or by rapid evolution due to the extensive use of insecticides in recent years.

Compositional diversity of the gut microbiota

Our gut microbiota sequencing of *H. cunea* yielded 8.65 GB of valid data after filtering of *H. cunea* genome sequences and produced 28,846,959 clean reads and 151,448 contigs with a total length of 520.68 Mb after de novo assembly (Table S8). Based on the alignment of sequencing reads to the NCBI RefSeq database, the microorganism composition was annotated (Table S9) and analyzed, and the microbes were grouped into taxonomic categories from kingdom to species level. We found 324 kingdoms, 135 phyla, 13 classes, 244 orders, 157 families, 200 genera, and 78 species in the larval gut of *H. cunea*.

At the phylum level, the *H. cunea* gut microbiota was dominated by Proteobacteria (71.33% of the total midgut bacteria contigs), followed by Euryarchaeota and Firmicutes (8.40 and 6.10% of the contigs, respectively) and to a lesser extent, Tenericutes, Actinobacteria, Cyanobacteria and Bacteroidetes; other phyla were less than 1% of the total contigs (Fig. 3a). At the class level, Gammaproteobacteria, Betaproteobacteria, Halobacteria and Clostridia comprised 77% of the contigs (Fig. 3b), while Enterobacteriales, Halobacteriales and Burkholderiales comprised 60% of all contigs at the order level (Fig. 3c). The three most abundant families were Enterobacteriaceae, Halobacteriaceae and Burkholderiaceae (50.86, 6.16, and 4.58% of total contigs, respectively) (Fig. 3d). At the genus level, microorganisms were rich in *Klebsiella*, *Halovivax* and *Burkholderia* (37.92, 4.75 and 4.32% of total contigs, respectively) (Fig. 3e). *Klebsiella oxytoca* was the most abundant species in the midgut of *H. cunea*, followed by *Halovivax ruber*, *Mannheimia haemolytica*, and *Burkholderia vietnamiensis* (Fig. 3f).

Functional annotation of the leaf-eating caterpillar gut metagenome

Our metagenomic analysis led to the identification of 102,787 nonredundant protein-coding genes with an average length of 300 bp (30.80 Mb total length) in the microbiota of the *H. cunea* larval gut. Gene functional annotation based on KEGG pathways showed that the most abundant function in the metagenome was metabolic function, representing 45.16% of all KEGG functions in the *H. cunea* gut microbiota.

KEGG iPath 2 analysis showed that the metabolic activities of the *H. cunea* gut bacteria were associated with digestion, nutrition and detoxification, including metabolism

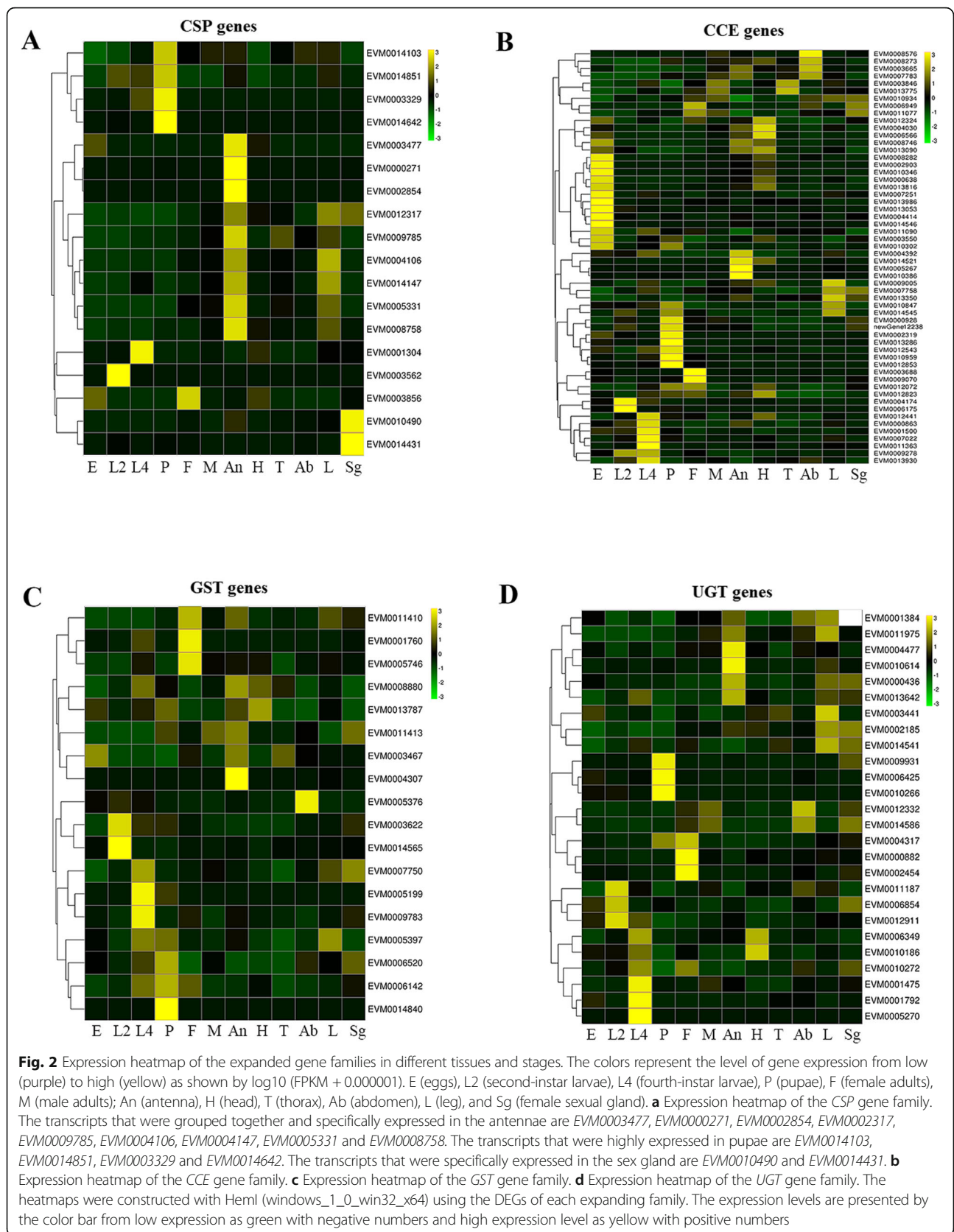
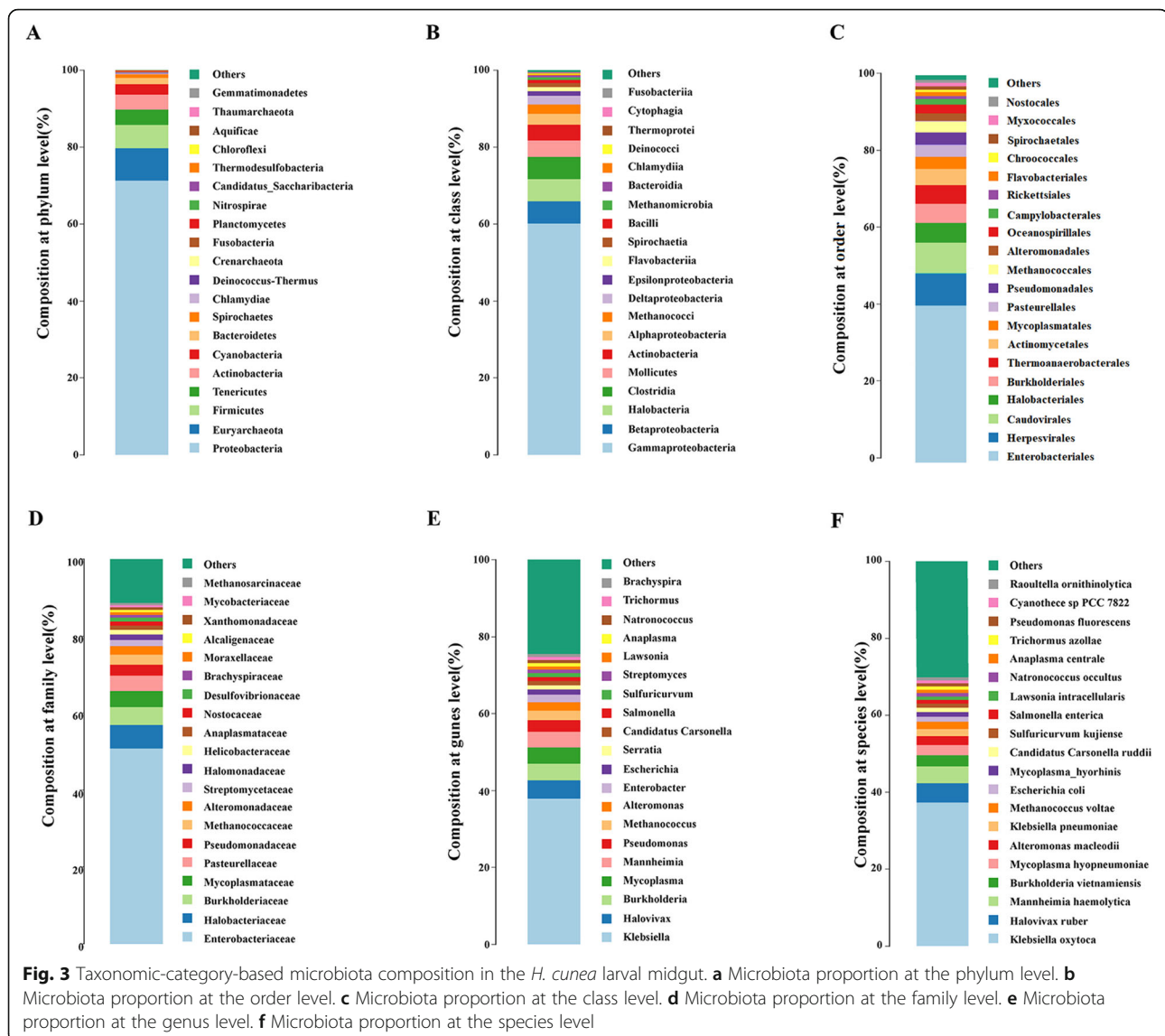


Table 4 Genes under significant positive selection (LRT, $p < 0.05$)

Gene ID	P-value	Annotation
EVM0013226	2.79E-11	ADP-ribosylation factor 6
EVM0009669	2.58E-09	RNA-binding protein
EVM0012652	4.88E-09	Proteasome non-ATPase regulatory subunit
EVM0001215	2.12E-07	WD repeat domain-containing protein 83
EVM0000300	3.76E-05	Adenylosuccinate lyase
EVM0012350	0.000269395	Calcium uptake protein 1
EVM0013607	0.000594247	SNF-related serine/threonine-protein kinase
EVM0005265	0.000734493	Glycerophosphocholine phosphodiesterase GPCPD1
EVM0001021	0.0008973	Flavin reductase (NADPH)
EVM0012819	0.001467642	D-aspartate oxidase
EVM0014230	0.001971819	Cyclin-Y-like protein 1
EVM0010560	0.002186741	Ras-related protein Rab-32
EVM0008629	0.00253065	Pyridine nucleotide-disulfide oxidoreductase domain-containing protein 1
EVM0004149	0.005305157	Pentatricopeptide repeat-containing protein 1, mitochondrial
EVM0004370	0.006201581	Replication factor C subunit 3
EVM0001595	0.00637425	WW domain-binding protein 2
EVM0004769	0.00712973	Heparan-sulfate 6-O-sulfotransferase 2
EVM0004055	0.00764959	Transmembrane protein 147
EVM0000132	0.01233362	Carbonic anhydrase 1
EVM0004959	0.01293744	Phosphatidylinositol glycan
EVM0000082	0.01403601	Disco-interacting protein 2
EVM0015138	0.05027551	Activin receptor type-1
EVM0003028	0.15142045	Cyclin A
EVM0013889	0.015549715	Phosphoserine aminotransferase
EVM0012220	0.015868726	Glycogen-binding subunit 76A
EVM0002611	0.016694366	E3 ubiquitin-protein ligase synoviolin A
EVM0005111	0.016757707	Prefoldin subunit 5
EVM0010778	0.018825687	Metaxin-1
EVM0006647	0.021983019	Vacuolar ATP synthase subunit E; V-type H ⁺ -transporting ATPase subunit E (A)
EVM0008673	0.024327862	Lysosome-associated membrane glycoprotein 1
EVM0000028	0.024646682	Transmembrane protein 183
EVM0010281	0.029467391	Secretion-regulating guanine nucleotide exchange factor-like
EVM0004884	0.030539908	Formin-binding protein 1-like
EVM0013117	0.034084621	Golgi apparatus protein 1-like
EVM0013726	0.03547635	Chitin binding domain protein
EVM0010443	0.036822765	3'(2'),5'-bisphosphate nucleotidase 1-like
EVM0010054	0.041270885	Very-long-chain enoyl-CoA reductase
EVM0009687	0.046760334	Cytochrome P450 306a1
EVM0002230	0.047306291	Ribosomal protein L27

of energy, carbohydrates, amino acids, lipids, cofactors, vitamins, glycans, xenobiotics, and terpenoids. The most enriched functions within these activities were “Folding, sorting and degradation”, representing 15.35% of all KEGG pathways, followed by “Signal transduction” (11.08%). The

nutrient metabolism functions that could be provided by gut microbiota were “Carbohydrate metabolism” (8.83%), “Amino acid metabolism” (7.09%), “Energy metabolism” (6.59%), “Nucleotide metabolism” (4.55%), “Lipid metabolism” (3.90%), “Glycan biosynthesis and metabolism”



(1.93%) and “Metabolism of cofactors and vitamins” (2.72%). In addition, genes in the gut microbiota were found with functions related to “Xenobiotics biodegradation and terpenoid metabolism” (2.09%) and “Biosynthesis of other secondary metabolites” (0.31%) (Figs. 4 and S9 and Table S10).

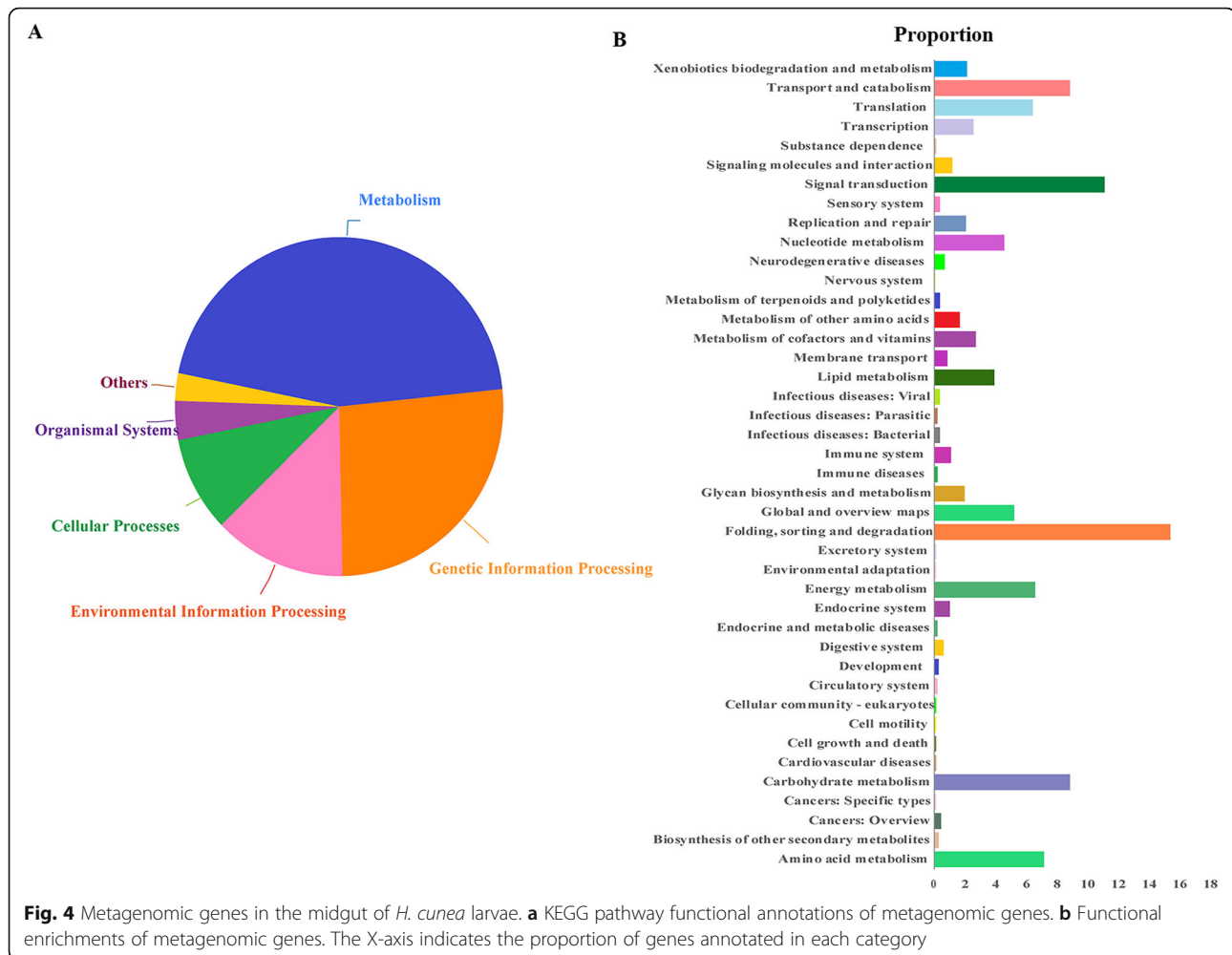
A total of 336 enzymes associated with cellulose and hemicellulose hydrolysis were identified in the intestinal flora of *H. cunea* based on the Carbohydrate-Active Enzyme (CAZy) database, including 42 auxiliary activities (AAs), 68 carbohydrate binding modules (CBMs), 75 carbohydrate esterases (CEs), 68 glycoside hydrolases (GHs), 82 glycosyltransferases (GTs) and one polysaccharide lyase (PL) (Figure S8 and Table S11). The results indicate that the gut microbes of *H. cunea* were most likely involved in cellulose degradation. By sequence alignment, we also predicted 55, 256 and 236 genes

possibly encoding for glutathione S-transferases, esterases, and P450s, respectively (Table S12).

Silk-web-related genes

Notably, one gene family related to silk production, Kazal-type serine proteinase inhibitors (*KSPIs*) [75], showed an expansion among the tested orthologous gene groups, which implies that silk-related genes might also have a role in the environmental adaptation to larval development of *H. cunea*. Hence, we performed further studies on silk-web-related genes.

The silk gland is a long paired organ of the fall web-worm. It specializes in the synthesis and secretion of silk proteins (Fig. 5a) and quickly atrophies after the onset of adulthood. The anatomy of the silk gland in the fall web-worm is quite similar to that of *B. mori*, and consists of three functionally distinct regions: the anterior silk gland

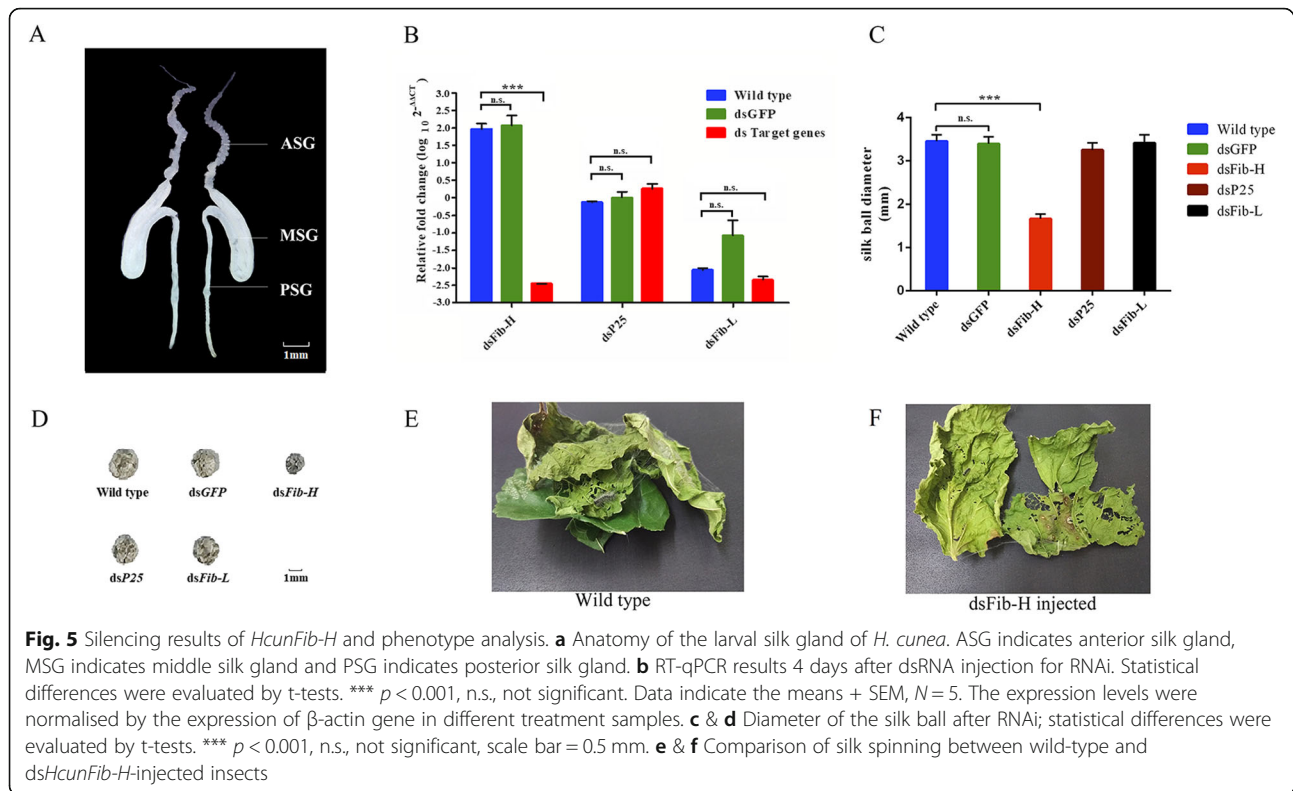


(ASG), middle silk gland (MSG) and posterior silk gland (PSG) [76]. Thirty-three silk-gland-related genes were identified in *H. cunea* (Table 5) through a homologous search against those from other Lepidopteran silk glands in previous studies [76–78], including 3 silk protein genes, 4 silk regulation genes and 26 protease inhibitor genes. In *B. mori*, the silk protein is composed of a ((Fib-H) - (Fib-L))₆ -P25 fibroin complex and held together by the protein sericin [79]. Here, three fibroin structure genes, *HcunFib-H*, *HcunFib-L* and *HcunP25*, were identified from the *H. cunea* genome, but our results showed that no sericin genes were annotated in the *H. cunea* genome; however, some silk regulation genes, such as silk gland factors (SGFs), fibroin-modulator-binding protein-1 (*FMBP-1*) and fibroinase, were identified in the *H. cunea* genome. Moreover, several protease inhibitors, such as kazal-type serine protease inhibitors, pacifastin-related serine protease inhibitor (pacifastin), phosphatidylethanolamine-binding protein, alpha-2-macroglobulin (A2M), cysteine proteinase inhibitor,

carboxypeptidase inhibitor, cystatin, serpins and proteasome inhibitor genes, were identified.

Silencing of silk fibroin genes and phenotype analysis

Because silk is a structural material and plays a crucial role in the survival of many insects, the extraordinary mechanical properties of silk are often explained in adaptive terms [80]. Fibroin is the key component of silk; it determines both the quantity and the structure of a silk web [81]. Here, three structural protein genes, *HcunFib-H*, *HcunFib-L* and *HcunP25*, were chosen for RNAi experiments to explore the mechanism of web production in *H. cunea* because of their involvement in silk production. We targeted three fibroin genes for silencing and measured their expression levels by qRT-PCR 4 days after injection (Fig. 5b). In comparison with the noninjected groups, there were no significant changes in the expression of *GFP*, *HcunFib-L* and *HcunP25* ($p > 0.05$), while the relative expression of *HcunFib-H* was dramatically decreased ($p < 0.001$). The different expression levels among the three genes might be one of reasons



that resulted in the differences in RNAi as the expression of *HcunFib-H* was much higher than those of *HcunFib-L* and *HcunP25*.

Within 10 days after the injection, the average diameters of the silk balls in the different treatments were as follows: that

of the noninjected wild type was 3.45 ± 0.12 mm; ds*GFP*-injected, 3.39 ± 0.14 mm ($p = 0.64 > 0.05$); ds*HcunFib-L*-injected, 3.41 ± 0.15 mm ($p = 0.79 > 0.05$); ds*HcunP25*-injected, 3.25 ± 0.14 mm ($p = 0.18 > 0.05$) and ds*HcunFib-H*-injected, 1.67 ± 0.09 mm ($p < 0.0001$)

Table 5 Identification of silk-web-related genes

Functional classification	Gene name	Gene ID
Silk protein genes	Fibroin light chain gene (<i>Fib-L</i>)	EVM0009358
	Fibroin heavy chain gene (<i>Fib-H</i>)	EVM0005282
	25 kDa silk protein (<i>P25</i>)	EVM0009847
Silk regulate genes	Sericin gene	None
	Fibroinase	EVM0000430
	Fibroin-modulator-binding protein-1	EVM0012647
Protease inhibitor	Silk gland factor	EVM0004972; EVM0012444
	Kazal-type serine protease inhibitor (<i>kazal</i>)	EVM0003147; EVM0009039; EVM0001538; EVM0000158; EVM0015003; EVM0006856; EVM0006205; EVM0013167; EVM0000098
	Pacifastin-related serine protease inhibitor (<i>pacifastin</i>)	EVM0013021
	Phosphatidylethanolamine-binding protein (<i>PBP</i>)	EVM0007560; EVM0006444; EVM0011908; EVM0008732; EVM0002652;
	Alpha-2-macroglobulin (<i>A2M</i>)	EVM0001565; EVM0002258
	Cysteine proteinase inhibitor	EVM0003884; EVM0006961; EVM0010793; EVM0002065; EVM0006398
	Carboxypeptidase inhibitor	EVM0003948
Cystatin	EVM0010793; EVM0012410	
Proteasome inhibitor	EVM0011839	

(Fig. 5c and d). There was a significant decrease in the quantity of silk in the ds*HcunFib-H* injected group, which was consistent with the dramatic decrease in the gene expression of *HcunFib-H* of the ds*HcunFib-H*-injected group after RNAi. The silencing of the silk structure protein gene *Fib-H* led to less silk production and damaged the leaf-silk shelter structure of the fall webworm by breaking the silk-leaf connections (Fig. 5e and f), suggesting that *HcunFib-H* contributes significantly to the formation of fibroin, to related web-producing behaviors and to the silk-web-related adaptations of *H. cunea*.

Discussion

In this study, the genome of the fall webworm we obtained was of high integrity by PacBio sequencing. And compared with other publicly available Lepidopteran genomes (*Plutella xylostella*, *Papilio polytes*, *Papilio machaon*, *Papilio xuthus*, *Pieris rapae*, *H. armigera*, *O. brumata*, *B. mori*, *S. frugiperda* and *P. bianor*), the *H. cunea* genome possesses a comparatively longer contig N50 (only smaller than the genome of *O. brumata*, *S. frugiperda* and *P. bianor*). The large genome size of *O. brumata* could be explained to a large extent by its higher repeat content, containing 53.5% repetitive elements in *O. brumata* genome (35.7% in *H. cunea* and 38.4% in *B. mori* genomes) [40]. However, the large genome of *H. cunea* is more likely to be caused by a larger average intron size, the mechanism is worthy of further study, because the average intron size of the *H. cunea* genome was 1491 bp, much larger than 1082 bp of *B. mori* and 139 bp of *O. brumata*. A similar phenomenon was also reported in the *Locusta migratoria* genome [82].

According to the result of the phylogeny of Lepidoptera, *H. cunea* and *H. armigera* were estimated to have diverged at the Eocene-Oligocene boundary, while from the late Eocene to early Oligocene, with the end of a continuous cooling event [83], deciduous trees that were better able to cope with large temperature changes began to overtake evergreen tropical species [84]. In North America, where *H. cunea* is native, litchi and cashew nut were the dominant trees in the early Oligocene [85]. With the expansion of temperate deciduous forests during this epoch, the food sources of the fall webworm increased, which might have contributed to the expansion of the host range of *H. cunea*.

The *CSP*, *CCE*, *GST* and *UGT* gene families were expanded in *H. cunea* compared to the tested Lepidopteran species, similar expansions of the chemosensory gene family have also been detected in other insect genomes [13, 18, 86, 87]. For example, studies of 22 mosquito species found that a distinct clade of *CSPs* was expanded in three *Culicinae* species [86], and a lineage-

specific expansion was present in the whitefly *Bemisia tabaci* compared with *Adelphocoris lineolatus*, *Aphis gossypii*, *Apolygus lucorum*, *Myzus persicae*, *Nilaparvata lugens* and *Sogatella furcifera* [18]. However, additional tests are still needed to determine whether these *CSPs* respond to host plant volatiles of *H. cunea*. Many studies have shown that *GSTs* and *CCEs* mediate insect tolerance to allelochemicals and contribute to resistance to a wide range of insecticides [88, 89], while *UGTs* play a crucial role in detoxification and in the regulation of xenobiotics in insects [90]. Previous studies of *Culex quinquefasciatus* and *Aedes aegypti* reported that the expansion of detoxification genes might involve in making these insects particularly adaptable to polluted water and contribute to their development of metabolic resistance to pyrethroid insect pesticides [91]. Therefore, the expansion of detoxification genes in *H. cunea* might also reflect their wide range of host plants thus adaptation by enhancing their capacity to detoxify xenobiotics and resist insecticides [92]. For other major expanded gene families, the hemolymph protein has been studied as an antifreeze protein, contributing to insect cold adaptation [93]; cecropin serves as an antibacterial protein of the insect immune system, supporting resistance to pathogenic microorganisms, and might be responsible for the adaptation of living organisms to environmental conditions [94]; serine proteases are known to dominate the lepidopteran larval gut environment and contribute to the polyphagous nature of insect pests such as *H. armigera* [95]. However, more experimental and field tests are needed to determine if the expansion of these three gene families could play important roles in the adaptation of *H. cunea* to environmental changes.

The nine *CSPs* were grouped together and specifically expressed in the antennae, strongly suggesting roles in olfactory sensing and host location. While the six *CSPs* expressed in pupae or sex gland might have other biological functions in *H. cunea* rather than those relating to olfaction such as carbon dioxide detection, larval development and leg regeneration reported previously [96]. For detoxification genes, in the lepidopteran *S. frugiperda*, detoxification genes such as *CCEs* and *GSTs* were much more highly expressed in lufenuron-resistant larvae than in lufenuron-susceptible larvae [97]. Some detoxifying genes in *Heliconius melpomene* larvae, including those encoding for *GSTs*, *UGTs* or *P450s*, responded significantly to a host plant shift [98]. In addition to lepidopteran insects, studies of other polyphagous species, *Tetranychus urticae* and *Anopheles gambiae*, also showed that the expression levels of many detoxification genes were significantly increased in association with host plant shifts or feeding stages [99, 100]. Furthermore, insect larvae are exposed to a range of food sources thus plant allelochemicals and a variety of

bacterial toxins. The ability of insect species to tolerate these toxins can influence their distribution [101]. These findings suggest that the detoxification genes that were highly expressed during the peak feeding period of *H. cunea* might contribute to *H. cunea* host plant adaptation or host shift [102]. Additionally, the positive selection of *HcunP450 CYP306A1* might reflect the rapid development of insecticide resistance in *H. cunea*, but much more testing is required.

There is growing evidence that the gut microbiota of insects plays crucial roles in diverse functions for the hosts, including growth, development and environmental adaptation [25]. Metagenomic approaches have been successfully applied to understand the relationship between gut microbiomes and their hosts over the past decade [103, 104]. The results of our study showed that the predominant phyla of bacteria in the gut of *H. cunea* larva were Proteobacteria and Firmicutes (Fig. 3a). Similar results have been found in previous studies of many different orders [25, 105]. These two phyla are also ubiquitous in the guts of lepidopteran insects such as *H. armigera* [106] and *B. mori* [107]. Proteobacterial symbionts are considered to be useful to the digestion of host insects [108] and to be involved in carbohydrate degradation and nitrogen fixation [109, 110], which help their hosts prevent the establishment and proliferation of pathogenic bacteria [111–114]. Firmicutes play a role in insecticide degradation and increase the abundance of resistant lines [26, 110, 112]. The three most abundant microbes in *H. cunea* gut at the genus level were *Klebsiella* (37.92%), *Halovivax* (4.75%) and *Burkholderia* (4.32%) (Fig. 3e); these results are very different from those of the gut microbiome of the host specialist *B. mori*. In the gut microbiota of *B. mori* (a standard inbred strain, Dazao), *Enterococcus* (18.7%), *Acinetobacter* (16.20%) and *Aeromonas* (8.70%) were the three most abundant microbial genera [115]. The genera *Klebsiella*, *Halovivax* and *Burkholderia* that occur in *H. cunea* have been reported to contribute to cellulose degradation [116], nitrogen fixation [117], carbon metabolism [118, 119], insect growth [120] and fenitrothion resistance [121]. Thus, the abundance of these bacteria might imply their contribution to host adaptation in *H. cunea* [122–124], but much more testing is required. Moreover, the carbon in plant cell walls exists in the form of cellulose, hemicelluloses, and lignin and is largely inaccessible to most organisms [125]. It is now well understood that gut symbiotic communities, most notably the symbiotic bacteria of termites [126] and ruminants [127], play a pivotal role in cellulose deconstruction in many invertebrates and vertebrates. Thus, the functional annotation of the leaf-eating caterpillar gut metagenome of the fall webworm was studied.

The gregarious larvae of *H. cunea* build conspicuous leaf-silk shelters on their host trees [128], where the larvae feed and live as they grow. The fall webworms generally aggregate in the web during daylight and extend their webs at night to enclose edible leaves for feeding [34, 129]. Moreover, this extended web can provide a protected space for larval development by blocking environmental damage or attacks from natural enemies, and is used as a support during ecdysis [130]. The silk web also regulates heat and slows air movement within the web, fostering the development of *H. cunea* larvae [34, 131]. The temperature inside the webs is considerably higher compared to the ambient temperature, and the interior heat-retention properties of the web rely mainly on the thickness, abundance and color of the web than on behavioral factors [34]. These physicochemical characteristics of the silk web are modified by the silk proteins [132]. Although, the anatomy of the silk gland of *H. cunea* is quite similar to that of *B. mori*, the composition of the silk protein between *H. cunea* and *B. mori* was different. In our case, there were no sericin genes identified from the genome and 12 transcriptome datasets of *H. cunea*. For sericin, some studies have shown that the sericin-related protease inhibitor in *B. mori* functions to protect the silk web or cocoon from degradation [133–137], but these glue-like proteins are absent in some spider species [138]. And some saturniid insects lack Fib-L and P25 proteins in the fibroin complex [130]. The silk contains 70%~75% fibroin [139], of which Fib-H accounts for 93% (w/w) of the composition in Lepidoptera [140]. To explore the function of the silk structure proteins in *H. cunea*. The three silk protein genes were selected from the 33 silk gland related genes to RNAi test. The result of RNAi suggesting that the *HcunFib-H* plays a critical role in the formation of fibroin, the web-producing behaviors and the silk-web-related adaptations of *H. cunea*. This gene could potentially be used as a target for future pest management of *H. cunea*.

For silk regulation genes, there is conclusive evidence showing that these genes are involved in regulating the synthesis or degradation of sericin and fibroin. *SGFs* stimulate the transcription of *sericin-1* via different binding sites [141, 142] and play a key role in regulating tissue-specific expression of the fibroin gene [143]. *EMBP-1* regulates the specificity of fibroin gene expression by binding the upstream and intronic promoter elements of the fibroin gene [144, 145]. Fibroinase in the silk gland is a cathepsin L-like cysteine proteinase that can digest silk proteins in the lumen of the silk gland after spinning and is regulated by protease inhibitors [146]. For protease inhibitors, many studies have shown that some protease inhibitors are specifically expressed in silk glands to avoid infection [77, 147]. The expanded

KSPI gene family, as a silk proteinase inhibitor 2 (SPI 2) in Lepidoptera, was reported to be involved in the inhibition of bacterial subtilizing and fungal proteinase K activity [75].

Conclusion

The first genome of the worldwide invasive pest *H. cunea* was obtained, and further studies revealed three causes of *H. cunea*'s adaptation. First, some chemosensory and detoxification gene families were expanded, suggesting the contribution of these genes to their extreme polyphagy at a genomic level. In addition, several nutrient metabolic and detoxification genes were found to evolve more rapidly along with their host expansions. Second, our results support that some gut microbes and their metabolic pathways are able to assist their nutrient metabolic and detoxification and might be involved in the host adaptation of *H. cunea*. Third, the silk web, which has been shown to function in the aggregated foraging and thermal regulation behavior of *H. cunea*, was further explored by silencing one of the silk protein gene *HcunFib-H*, significantly decreasing the quantity of silk and breaking silk-leaf connections. Overall, our results provide some evidence on the adaptation of *H. cunea*, partially explaining the reasons for the rapid invasion of *H. cunea* at the genome, transcriptome and metagenome levels, along with some potential gene targets and innovative strategies for the control of this invasive pest.

Methods

Insects

A colony of *H. cunea* was established from a single egg mass and maintained in our laboratory for population expansion to reduce heterozygosity. Because low genomic heterozygosity is important for obtaining high-quality genomes, we used a fifteen-generation inbred population of *H. cunea* from a single egg mass. The egg mass was collected in the field from damaged forest in Beiling Park, Liaoning Province, China, in 2015. The colony was fed fresh mulberry leaves (26 °C, 80% RH, 19:5 light:dark cycle, BIC-300 artificial climate chest, Boxun, Shanghai, China). Genomic DNA was extracted using the cetyltrimethylammonium bromide (CTAB) method from a single adult male. The sample was washed with double-distilled water and frozen in liquid nitrogen before DNA extraction. After measuring the concentration and quality, the genomic DNA was immediately stored in a -80 °C freezer until further sequencing.

Genome sequencing and assembly

First, we performed a preliminary survey to evaluate the genome size, repeat sequence ratio and heterozygosity of the *H. cunea* genome; for this, a genome survey with k-mer analysis was used as a general and assembly-

independent method for estimating these three genomic characteristics as mentioned above, the 270 bp library data were used to construct a k-mer distribution map for $k = 19$. Then, genome sequencing was performed on two separate platforms (PacBio and Illumina). For PacBio genome sequencing, the genomic DNA was sheared using g-TUBE devices (Covaris, Inc., USA) and purified using a 0.45 volume ratio of AMPure PB beads. SMRTbell libraries were created using the 'Procedure & Checklist—20 kb Template Preparation using BluePippin™ Size Selection' protocol [148]. The quality of the library was tested with a Qubit fluorometer (Invitrogen Life Technologies, CA, USA) and an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). Then, the library was sequenced on a PacBio RSII (Expression Analysis, Durham, NC, USA) platform. For Illumina sequencing, two libraries with insert sizes of 270 bp and 500 bp were built with the Ultra II DNA Library Prep Kit (New England Biolabs, Ipswich, MA, USA) and sequenced by using the Illumina HiSeq X Ten system (Illumina, San Diego, CA, USA) with rapid runs. All sequencing was conducted by Biomarker Technology Co., Ltd. (Beijing, China).

The sequencing adapters and the low-quality reads (with read quality score < 50 or length < 50 bp) in the sequencing reads were removed by the RS Subreads Protocol of SMRT Analysis version 2.3 [149]. We corrected the PacBio reads with Canu version 1.7 [150] and then assembled the retained high-quality subreads with Canu v1.5, Falcon v0.7 and WTDBG v1.2.8 [151, 152] independently. Finally, the draft assembly was corrected and polished with Pilon [153] using high-coverage Illumina reads. Based on the optimal assembly results, we evaluated the completeness of the genome assembled by WTDBG. The alignment efficiencies were calculated by mapping the reads generated by the Illumina platform and the reads corrected by Canu to the assembled genome. Then, two databases, the Core Eukaryotic Genes Mapping Approach (CEGMA v2.5) [154] and Benchmarking Universal Single-Copy Orthologs (BUSCO v2.0) [155], were used to assess the completeness of the WTDBG assembly.

Repeats and noncoding RNAs

The specific repetitive sequence database was used to predict repeat sequences. A de novo repeat library of *H. cunea* was constructed by LTR_FINDER v1.05 [156], MITE-Hunter [157], RepeatScout v1.0.5 [158] and PILER-DF v2.4 [159]; then, it was classified by PASTE Classifier [160] and combined with the Repbase transposable element library to act as the final library. Afterward, RepeatMasker v4.0.6 [161] was used to find the homologous repeats in the final library. tRNAscan-SE v1.3.1 [162] was used to search

for tRNA coding sequences. rRNA and microRNA were identified by Infernal v1.1 [163] based on the Rfam database and miRBase database. The pseudogene were predicted by two steps: Firstly, GenBlastA v1.0.4 [164] was applied to identify the candidate pseudogene by homologous searching against genome data. Secondly, GeneWise v2.4.1 [165] was performed to search for immature termination and frameshift mutation of pseudogene .

Gene prediction and functional annotation

To identify protein-coding sequences, a combination of ab initio gene prediction, homology-based prediction and unigene-based methods were used as annotation pipelines. Genscan [166], Augustus v2.4 [167], GlimmerHMM v3.0.4 [168], GeneID v1.4 [169] and SNAP [170] were used to predict the protein-coding sequences. Four species (*Amyeloid transitella*, *Bombyx mori*, *Helicoverpa armigera* and *Plutella xylostella*) were used to complete homology-based gene prediction with GeMoMa v1.3.1 [171]. Reference transcriptome assembly was performed by HISAT v2.0.4 and StringTie [172], and gene prediction was performed by TransDecoder v2.0 [173] and GeneMarkS-T v5.1 [174]. The de novo transcriptome was completed by PASA v2.0.2 [175]. Finally, all the results from three gene prediction methods (GeMoMa, TransDecoder v2.0 and GeneMarkS-T) were integrated by EvidenceModeler (EVM) v1.1.1 [176] and annotated by PASA v2.0.2.

Gene functions were assigned according to the best-match BLASTp alignments in the NR databases, KOG, TrEMBL and Kyoto Encyclopedia of Genes and Genomes (KEGG). GO annotations were obtained by Blast2GO based on the results of alignment to NR. Moreover, we also performed enrichment analyses of the Clusters of Orthologous Groups of proteins (COG), GO terms and KEGG pathways.

Orthologous gene families

The most updated genome sequences of twelve sequenced insects (*Apis mellifera*, *Bombyx mori*, *Drosophila melanogaster*, *Helicoverpa armigera*, *Papilio machaon*, *Papilio polytes*, *Papilio xuthus*, *Pieris rapae*, *Plutella xylostella*, *Tribolium castaneum*, *Hyphantria cunea*, and *Operophtera brumata*) were used to infer gene orthology and construct the phylogenetic tree, the details of these genomic datasets we used in this study were shown in Table S13. After downloading the annotated coding sequences from NCBI, the longest protein sequences per gene were extracted to perform a best reciprocal hit (BRH) analysis by all-v-all BLAST using an E-value equal to $1E-05$ to identify orthologous genes among the twelve species by OrthoMCL 5 [177].

Phylogenetic tree and divergence times

The longest open reading frames (ORFs) for the longest transcript pairs across the twelve species were extracted by a Perl script, and tORFs in each orthologous set were aligned using PRANK [178] with the following parameters: $-f = \text{fasta}$ $-F = \text{codon}$ $-\text{noxml}$ $-\text{notree}$ $-\text{nopost}$. The alignment for each locus was trimmed by Gblocks v 0.91b [179] (Parameters: $-\text{t} = \text{c}$ $-\text{b3} = 1$ $-\text{b4} = 6$ $-\text{b5} = \text{n}$) to reduce the rate of false positive predictions by filtering out sequencing errors, incorrect alignments and non-orthologous regions based on codons [180]. After trimming, alignments of less than 120 bp were removed. The single-copy orthologous genes were concatenated into one supergene, and the best amino acid substitution model was estimated. RAXML v. 8.0.26 [181] was used to construct the phylogenetic tree based on the supergene under the LG + I + G + F model with 1000 bootstrap replicates. The divergence times among species were estimated by R8s v. 1.7.1 [182] with a node dating approach that used three fossil records as the most recent common ancestor. The three fossil records we used in this study were the oldest definitive beetle (*Coleopsis archaica* gen. et sp. Nov., 298.9 to 295.0 Ma), the oldest fossil Diptera (such as: *Anisnodus crinitus* n. gen., n. sp., 247.2 to 242.0 Ma) and the oldest fossil Rhopalocera (*Praepapilio Colorado* n. g., n. sp., *P. gracilis* n. sp., and *Praepapilioninae. Riodinella nympa* n. g., n. sp., 46.2 to 40.4 Ma), respectively [183–185].

Gene family expansion/contraction

CAFE [186] was used to examine the expansion and contraction of gene families among the twelve species. The results of the orthologous gene identification were filtered by CAFE's built-in script, and the global parameter λ was estimated by the maximum likelihood method. Comparing divergence size and species size calculated by CAFE could determine whether expansion had occurred. The divergence size indicates the ancestral gene family size for each node in the phylogenetic tree, and the species size indicates the gene number in the homologous gene family. When the divergence size is smaller than the species size, the gene family is expanding. Additionally, for each gene family, a conditional *P*-value was calculated, and gene families with *P*-values < 0.05 were considered to have significantly expanded or contracted.

Positive selection analyses

A branch-site model (parameters: Null hypothesis: model = 2, NSsites = 2, fix_omega = 1, omega = 1; alternative hypothesis: model = 2, NSsites = 2, fix_omega = 0, omega = 1) in PAML [187] was used to identify the genes with positively selected sites in the fall webworm genome using our tree topology as the guide tree. Then,

likelihood ratio tests (LRTs) were performed to detect positive selection on the foreground branch. Only those genes with LRT *P*-values less than 0.05 were inferred as positively selected.

Transcriptome analysis of different stages and tissues

RNA sequencing was performed on different developmental stages and tissues of *H. cunea*. The following developmental stages were selected for the transcriptome analyses: eggs, second instar larvae, fourth instar larvae, pupae, and male and female adults. The following tissues were used for the tissue transcriptome experiment: head, thorax, leg, abdomen, antenna, and female sexual glands. For each group, fifteen individuals were mixed for RNA extraction, and three biological replicates were produced for each sample. Total RNA was isolated from the homogenized samples using TRIzol reagent (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's protocols. After extraction, total RNA was assessed with the NanoDrop 2000 (Thermo Fisher Scientific, Waltham, MA, USA) and the Agilent Bioanalyzer 2100 System (Agilent Technologies, CA, USA) to verify the integrity and quality of RNA.

After each sample was quantified, the libraries were built and sequenced on the Illumina HiSeq X Ten platform. After filtering, clean reads were mapped to the reference genome sequence obtained in this study with Hisat2 tools [188]. Only reads with a perfect match or one mismatch were retained for further analysis. Cufflinks counts the expression of each gene and reports it in fragments per kilobase of transcript per million fragments mapped (FPKM) [189]. For each sequenced library, the read counts were adjusted by the edgeR package [190] with one scaling normalized factor. Differentially expressed gene (DEG) analysis within two sample groups (stages and tissues) was performed using the EBSeq R package [191], and then the false discovery rate (FDR) was performed based on the Benjamini-Hochberg (BH) procedure [192] to correct the *P* value of the identified datasets, with the standard of $FDR \leq 0.01$ and fold change (FC) ≥ 2 to remove the false positive datasets.

Metagenomic sequencing and analysis

To test whether symbiotic microbes facilitate environmental adaptation in *H. cunea*, detailed profiles of the gut microflora were obtained by metagenomic sequencing. Midgut samples were collected from ten last-instar larvae of *H. cunea* from the same wild population on their host plant (*Quercus mongolica*) and preserved in RNAlater. DNA was extracted from a mixture of ten gut samples using an effective gut microbiota DNA extraction kit (QIAamp DNA Stool Mini Kit; Qiagen) and stored at -20°C . A paired-end gut microbiota DNA library was built using the NEBNext DNA Library Prep

Mast Mix Set for Illumina (New England Biolabs, Ipswich, MA, USA). Sequencing was then performed on the Illumina HiSeq platform. The raw reads were checked and filtered by the following methods: 1) reads with adapters were removed; 2) reads with low-quality and N bases (quality value ≤ 10) were removed; 3) to gain a clearer understanding of the bacterial genome data, the host genome data were filtered out by eliminating fall webworm genome sequences.

Kraken [193] was used for the taxonomic identification and relative abundance calculations, and the NCBI Reference Sequence Database (RefSeq), which includes high-quality bacterial, archaea and virus data, could further filter the nonbacterial genome sequences. The microbiota composition was visualized by Krona [194] and Python scripts.

De novo assembly was performed by IDBA-UD [195] (parameter: `--mink:21, --maxk:101, --step:20, --pre_correction`), resulting in sequences greater than 500 bp. The assembly quality was assessed by QUASt [196]. MetaGeneMark [197] was used to perform ab initio gene prediction with the default settings. Prophage prediction was performed by BLAST (E-value: $1\text{E-}05$) with a local database based on the ACLAME database. Transposable elements, including DNA transposons, long terminal repeats (LTRs), long interspersed elements (LINEs) and short interspersed elements (SINEs), were identified by using RepeatMasker v 4.0.5 and RepeatProteinMasker [161]. A nonredundant data set was outputted by CD-HIT [198] with a minimum coverage cut-off of 0.9 for the shorter sequences. All genes in our nonredundant dataset were translated into amino acid sequences and aligned to relevant databases: NR, COG, KEGG, Swiss-Prot, CAZy and ARDB by BLASTP (E-value $\leq 1\text{E-}05$). Blast2GO was used to obtain GO annotations, and HMMER v 3.0 [199] was used to annotate sequences in our dataset from the Pfam database.

RNA interference with silk fibroin genes

To study the web-producing mechanism in the fall webworm, silk-gland-related genes were identified by analyzing the *H. cunea* genome and the silk gland transcriptome produced in this study. Three genes encoding structural proteins, fibroin heavy (*Fib-H*), fibroin light (*Fib-L*) and protein 25 (*P25*), were silenced by RNA interference (RNAi) to examine their biological functions. RNAi was performed by injecting the corresponding gene-specific double-stranded RNAs (dsRNAs), and green fluorescent protein (*GFP*) was used as a negative control. The dsRNAs for *HcunFib-H*, *HcunFib-L*, *HcunP25* and *GFP* were synthesized by using the MEGAscript RNAi Kit (Ambion, Austin, TX, USA) following the manufacturer's procedure and purified by lithium chloride precipitation. After quantification with

a NanoDrop 2000 (Thermo Fisher Scientific, Wilmington DE, USA) and 1% agarose gel electrophoresis, the dsRNA of the four genes was stored at -80°C before use. Then, newly molted third-instar larvae were injected with $4\ \mu\text{g}$ of targeted dsRNA in $1\ \mu\text{L}$ into the abdomen using a Nanoliter 2000 injector (World Precision Instruments, Sarasota, FL, USA). In total, 20 individuals were injected, divided into four plastic boxes ($20\ \text{cm} \times 10\ \text{cm} \times 5\ \text{cm}$) and fed fresh mulberry leaves ($6\ \text{g}$ per box per day); of these, 15 individuals were used to observe the phenotype ($N = 3$), while 5 individuals were used for RT-qPCR validation ($N = 5$). The effect of RNAi was examined by RT-qPCR 4 days after injection; each cDNA sample was quantified based on the total RNA ($2\ \mu\text{g}$) from the 5 insects separately before reverse transcription (SuperScript™ III First-Strand Synthesis SuperMix), and β -actin was employed as an internal control. RT-qPCR was performed on a StepOnePlus Real-Time PCR Detection System (Bio-Rad, Hercules, CA, USA) using TransStar Tip Top Green qPCR Supermix (TransGen Biotech, Beijing, China). The silk web was collected from each box within 10 days after injection. Because the silk filaments were difficult to quantify, they were rolled into a tight ball, and the diameter of the silk ball was used to calculate the silk quantity. The RT-qPCR data were analyzed by the $2^{-\Delta\Delta\text{CT}}$ method. The primers used in this study are listed in Table S14, and the efficiency of each primer pair was tested before the RT-qPCR experiments.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s12864-020-6629-6>.

Additional file 1: Figure S1. K-mer distribution of preprocessed data with $k = 19$. The distribution of depth analysis based on whole genome data in the fall webworm. Using the formula: genome size = $k\text{-mer count/peak}$ of the kmer distribution, thereinto, $k = 19$.

Additional file 2: Figure S2. Numbers of genes annotated with three gene prediction strategies. The final number of genes supported by homologous prediction and transcriptome prediction was 14,688, accounting for a significant proportion (95.88%) of 15,319 (the total number of protein-coding genes), showing the high quality of the prediction.

Additional file 3: Figure S3. GO annotation of the *H. cunea* genome. The capital letters on the x-axis indicate the GO categories as listed below, the left y-axis indicates the percentage of genes in each category, and the right y-axis indicates the number of genes in each category.

Additional file 4: Figure S4. KOG annotation of the *H. cunea* genome. The capital letters on the x-axis indicate the KOG classification as listed on the right, and the y-axis indicates the number of genes in each classification.

Additional file 5: Figure S5. The distribution of Nr homologous genes within the *H. cunea* genome in insect species. The percentage of Nr homologous genes over the *H. cunea* genome were obtained by EvidenceModeler (EVM) with more ten insect species, including *Bombyx mori*, *Danaus plexippus*, *Helicoverpa armigera*, *Papilio xuthus*, *Manduca sexta*, *Spodoptera frugiperda*, *Papilio polytes*, *Tribolium castaneum*, *Spodoptera litura* and *Spodoptera exigua*.

Additional file 6 Figure S6. Box plot of FPKM values from different developmental stages and tissues. Box plot of \log_{10} FPKM values aggregated across the 8232 DEGs of the stage RNA sequencing groups and 7733 DEGs of the tissue RNA sequencing groups.

Additional file 7: Figure S7. Numbers of alternative splicing events in different tissues and stages of *H. cunea*. The horizontal axis represents the number of alternative splicing events under the corresponding event, and the vertical axis represents the abbreviation of the classification of alternative splicing events. (1) AE: Alternative exon ends; (2) IR: Intron retention (IR_ON, IR_OFF pair); (3) MIR: Multi-IR (MIR_ON, MIR_OFF pair); (4) MSKIP: Multiexon SKIP (MSKIP_ON, MSKIP_OFF pair); (5) SKIP: Skipped exon (SKIP_ON, SKIP_OFF pair); (6) TSS: Alternative 5' first exon (transcription start site); (7) TTS: Alternative 3' last exon (transcription terminal site); (8) XAE: Approximate AE; (9) XIR: Approximate IR (XIR_ON, XIR_OFF pair); (10) XMIR: Approximate MIR (XMIR_ON, XMIR_OFF pair); (11) XMSKIP: Approximate MSKIP (XMSKIP_ON, XMSKIP_OFF pair); (12) XSKIP: Approximate SKIP (XSKIP_ON, XSKIP_OFF pair).

Additional file 8: Figure S8. COG functional enrichment analysis annotation of the metagenomic data. L, R, C, G and O are the top five gene function categories. The remaining categories were defined as "Others" which including N: cell motility; D: Cell cycle control, cell division, chromosome partitioning; A: RNA processing and modification; M: Cell wall/membrance/envelope biogenesis; U: Intracellular trafficking, secretion, and vesicular transport.

Additional file 9: Figure S9. Proportion of Carbohydrate-Active Enzymes in the metagenome data. BLASTp was used to compare the sequences of the nonredundant gene sets with the CAZY database to obtain the gene annotation information.

Additional file 10: Table S1. Results of the preliminary survey. **Table S2.** Prediction of noncoding genes. **Table S3.** Statistical information of two-algebra data by Burrow-Wheeler Aligner (BWA). **Table S4.** Integrity evaluation of Illumina data. **Table S5.** Integrity evaluation of PacBio data by CEGMA. **Table S6.** Integrity evaluation of PacBio data by BUSCO. **Table S7.** DEG statistical results from different stages and tissues. **Table S8.** Sequencing statistics of metagenomic data. **Table S13.** The detail of the genome versions used in this study. **Table S14.** Primers used for RNA interference and RT-qPCR.

Additional file 11: Table S9. Annotation results of the metagenome data. **Table S10.** KEGG enrichment analysis of metagenomic genes. **Table S11.** Metagenomic enzymes associated with cellulose and hemicellulose hydrolysis. **Table S12.** Metagenomic gene sets encoding glutathione S-transferase, esterases, and *P450s*.

Abbreviations

A2M: Alpha-2-macroglobulin; AAs: Auxiliary activities; ABCs: ATP-binding cassette transporters; ASG: Anterior silk gland; CAZY: Carbohydrate-Active enZymes; CBMs: Carbohydrate binding modules; CCEs: Carboxyl/choline esterases; CEs: Carbohydrate esterases; COG: Cluster of Orthologous Groups; CSPs: Chemosensory Proteins; DEGs: Differential expression genes; dsRNAs: double-stranded RNAs; Fib-H: Fibroin heavy chain gene; Fib-L: Fibroin light chain gene; FMBP-1: Fibroin-modulator-binding protein-1; FPKM: Kilobase of transcript per million fragments mapped; GFP: Green fluorescent protein; GHs: Glycoside hydrolases; GRs: Gustatory Receptors; GSTs: Glutathione S-transferases; GTs: Glycosyltransferases; KEGG: Kyoto Encyclopedia of Genes and Genomes; KOG: euKaryotic Orthologous Groups; KSPIs: Kazal-type serine proteinase inhibitor; LINE: Long Interspersed Elements; LRTs: Likelihood ratio tests; LTR: Long Terminal Repeat; MSG: Middle silk gland; NR: NCBI non-redundant; OBPs: Odorant Binding Proteins; ORs: Odorant Receptors; P25: Protein 25; Pacifastin: Pacifastin-related serine protease inhibitor; PBP: Phosphatidylethanolamine-binding protein; Pfam: Protein family; PLs: Polysaccharide lyases; PSG: Posterior silk gland; RefSeq: Reference Sequence Database; SGFs: Silk gland factors; SINE: Short Interspersed Elements; SPI 2: Silk proteinase inhibitor 2; UGTs: UDP glycosyltransferases

Acknowledgements

We appreciate the critical editing and proofreading of JJ Scientific Consultant Ltd., UK, and American Journal Experts.

Authors' contributions

Qi Chen and Hanbo Zhao contributed equally to this work. YLW and BZR designed the project; Biomarker Technologies Corporation performed the sequencing, assembly and functional annotation work; QC, HBZ and JXL performed the data analyses; JXL, JTW, MW, HFZ and YXZ performed the experiments; QC, HBZ, YLW and JJZ wrote the manuscript with input from all the other authors, and all authors read and approved the final manuscript version.

Funding

Funding for this study came from the National Natural Science Foundation of China (No. 31501890, YLW), the Natural Science Foundation of Jilin Province (No. 20160204023NY and No. 20170204003NY, BZR; No. 20180101005JC and No. 20190301047NY, YLW), the Foundation of Xinjiang Uygur Autonomous Region (No. 2017E0272), the Open Project Program of the Jilin Provincial Key Laboratory of Animal Resource Conservation and Utilization (No. 130028684 and No. 1300289103), the Fundamental Research Funds for the Central Universities (No. 11SSXT153, No. 2412015KJ015 and No. 2412019FZ022), the Fund for Fostering Talents in Basic Science of the National Natural Science (No. J1210070), and the Undergraduate teaching quality and teaching reform project of Northeast Normal University (No. 131004003). The funding bodies played no role in writing the manuscript.

Availability of data and materials

The genome data of *H. cunea* have been deposited in the SRA under the accession number SUB5033887.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Jilin Provincial Key Laboratory of Animal Resource Conservation and Utilization, Northeast Normal University, Changchun, Jilin, China. ²Key Laboratory of Vegetation Ecology, MOE, Northeast Normal University, Changchun, China. ³School of Life Sciences, Changchun Normal University, Changchun, Jilin, China. ⁴Meihekou Forest Pest Control Station, Changchun, Jilin, China. ⁵Garden and Plant Protection Station of Changchun, Changchun, Jilin, China. ⁶Rothamsted Research, Harpenden AL5 2JQ, UK.

Received: 21 August 2019 Accepted: 26 February 2020

Published online: 18 March 2020

References

- Schowalter T, Ring D. Biology and Management of the Fall Webworm, *Hyphantria cunea* (Lepidoptera: Erebidae). *J Integr Pest Manage*. 2017;8(1):7.
- Ge X, He S, Zhu C, Wang T, Xu Z, Zong S. Projecting the current and future potential global distribution of *Hyphantria cunea* (Lepidoptera: Arctiidae) using CLIMEX[J]. *Pest Manag Sci*. 2019;75(1):160-69.
- Cocquemot C, Lindelöw A. BIORISK-biodiversity and ecosystem risk assessment, vol. 4. Sofia: Pensoft Publishers; 2010. p. 193–218.
- Sullivan GT, Karaca I, Ozman-Sullivan SK, Kara K. Tachinid (Diptera: Tachinidae) parasitoids of overwintered *Hyphantria cunea* (Drury)(Lepidoptera: Arctiidae) pupae in hazelnut plantations in Samsun province, Turkey. *J Ent Res Soc*. 2012;14:21–30.
- Chapman R. Chemosensory regulation of feeding. Regulatory mechanisms in insect feeding: Springer; 1995. p. 101–36.
- Qin J, Wang C. The relation of interaction between insects and plants to evolution[J]. *Acta Ecol Sin*. 2001;44(3):360-65.
- Ishikawa S, Hirao T, Arai N. Chemosensory basis of hostplant selection in the silkworm. *Entomologia Experimentalis et Applicata*. 1969;12(5):544–54.
- CORBET SA. Insect chemosensory responses: a chemical legacy hypothesis. *Ecol Entomol*. 1985;10(2):143–53.
- Sánchez-Gracia A, Vieira F, Rozas J. Molecular evolution of the major chemosensory gene families in insects. *Heredity*. 2009;103(3):208.
- Leal WS. Odorant reception in insects: roles of receptors, binding proteins, and degrading enzymes. *Annu Rev Entomol*. 2013;58(1):373–91.
- Benton R. Multigene family evolution: perspectives from insect chemoreceptors. *Trends Ecol Evol*. 2015;30(10):590–600.
- Simon J-C, d'Alençon E, Guy E, Jacquin-Joly E, Jaquiere J, Nouhaud P, et al. Genomics of adaptation to host-plants in herbivorous insects. *Brief Function Genomics*. 2015;14(6):413–23.
- Gouin A, Bretaudeau A, Nam K, Gimenez S, Aury JM, Duvic B, et al. Two genomes of highly polyphagous lepidopteran pests (Spodoptera frugiperda, Noctuidae) with different host-plant ranges. *Sci Rep*. 2017;7(1):11816.
- Robertson HM, Wanner KW. The chemoreceptor superfamily in the honey bee, *Apis mellifera*: expansion of the odorant, but not gustatory, receptor family. *Genome Res*. 2006;16(11):1395–403.
- Wanner K, Robertson H. The gustatory receptor family in the silkworm moth *Bombyx mori* is characterized by a large expansion of a single lineage of putative bitter receptors. *Insect Mol Biol*. 2008;17(6):621–9.
- Obiero GF, Mireji PO, Nyanjom SR, Christoffels A, Robertson HM, Masiga DK. Odorant and gustatory receptors in the tsetse fly *Glossina morsitans morsitans*. *PLoS Negl Trop Dis*. 2014;8(4):e2663.
- Opachaloemphan C, Yan H, Leibholz A, Desplan C, Reinberg D. Recent advances in behavioral (Epi) genetics in Eusocial insects. *Annu Rev Genet*. 2018;52:489–510.
- Zeng Y, Yang YT, Wu QJ, Wang SL, Xie W, Zhang YJ. Genome-wide analysis of odorant-binding proteins and chemosensory proteins in the sweet potato whitefly, *Bemisia tabaci*. *Insect Sci*. 2019;26(4):620–34.
- Despres L, David J-P, Gallet C. The evolutionary ecology of insect resistance to plant chemicals. *Trends Ecol Evol*. 2007;22(6):298–307.
- Edger PP, Heidel-Fischer HM, Bekaert M, Rota J, Glöckner G, Platts AE, et al. The butterfly plant arms-race escalated by gene and genome duplications. *Proc Natl Acad Sci*. 2015;112(27):8362–6.
- Rane RV, Walsh TK, Pearce SL, Jermin LS, Gordon KH, Richards S, et al. Are feeding preferences and insecticide resistance associated with the size of detoxifying enzyme families in insect herbivores? *Curr Opin Insect Sci*. 2016;13:70–6.
- Hidaka T. Adaptation and speciation in the fall webworm; 1977.
- Rajagopal R. Beneficial interactions between insects and gut bacteria. *Indian J Microbiol*. 2009;49(2):114–9.
- Engel P, Moran NA. The gut microbiota of insects—diversity in structure and function. *FEMS Microbiol Rev*. 2013;37(5):699–735.
- Krishnan M, Bharathiraja C, Pandiarajan J, Prasanna VA, Rajendhran J, Gunasekaran P. Insect gut microbiome—An unexploited reserve for biotechnological application. *Asian Pac J Trop Biomed*. 2014;4:516–21.
- Dillon R, Dillon V. The gut bacteria of insects: nonpathogenic interactions. *Annu Rev Entomol*. 2004;49(1):71–92.
- Dillon R, Charnley K. Mutualism between the desert locust *Schistocerca gregaria* and its gut microbiota. *Res Microbiol*. 2002;153(8):503–9.
- Wernegreen JJ. Mutualism meltdown in insects: bacteria constrain thermal adaptation. *Curr Opin Microbiol*. 2012;15(3):255–62.
- Mueller UG, Mikheyev AS, Hong E, Sen R, Warren DL, Solomon SE, et al. Evolution of cold-tolerant fungal symbionts permits winter fungiculture by leafcutter ants at the northern frontier of a tropical ant–fungus symbiosis. *Proc Natl Acad Sci*. 2011;108(10):4053–6.
- Montllor CB, Maxmen A, Purcell AH. Facultative bacterial endosymbionts benefit pea aphids *Acyrtosiphon pisum* under heat stress. *Ecol Entomol*. 2002;27(2):189–95.
- Russell JA, Moran NA. Costs and benefits of symbiont infection in aphids: variation among symbionts and across temperatures. *Proc R Soc B Biol Sci*. 2005;273(1586):603–10.
- Perlman SJ, Kelly SE, Hunter MS. Population biology of cytoplasmic incompatibility: maintenance and spread of *Cardinium* symbionts in a parasitic wasp. *Genetics*. 2008;178(2):1003–11.
- Loewy KJ, Flansburg AL, Grenis K, Kjeldgaard MK, Mccarty J, Montesano L, et al. Life history traits and rearing techniques for fall webworms (*Hyphantria cunea* Drury) in Colorado. *J Lepidopterists' Soc*. 2013;67(3):196–205.
- Rehnberg BG. Heat retention by webs of the fall webworm *Hyphantria cunea* (Lepidoptera: Arctiidae): infrared warming and forced convective cooling. *J Therm Biol*. 2002;27(6):525–30.
- Mondal M. The silk proteins, sericin and fibroin in silkworm, *Bombyx mori* Linn.,-a review. *Caspian J Environ Sci*. 2007;5(2):63–76.
- Devi R, Deori M, Devi D. Evaluation of antioxidant activities of silk protein sericin secreted by silkworm *Antheraea assamensis* (Lepidoptera: Saturniidae). *J Pharm Res*. 2011;4(12):4688–91.

37. Zurovec M, Kludkiewicz B, Fedic R, Sulitkova J, Mach V, Kucerova L, et al. Functional conservation and structural diversification of silk sericins in two moth species. *Biomacromolecules*. 2013;14(6):1859–66.
38. Stewart RJ, Wang CS. Adaptation of caddisfly larval silks to aquatic habitats by phosphorylation of H-fibroin serines. *Biomacromolecules*. 2010;11(4):969–74.
39. Xia Q, Zhou Z, Lu C, Cheng D, Dai F, Li B, et al. A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science*. 2004;306(5703):1937–40.
40. Derks MF, Smit S, Salis L, Schijlen E, Bossers A, Mateman C, et al. The genome of winter moth (*Operophtera brumata*) provides a genomic perspective on sexual dimorphism and phenology. *Genome Biol Evol*. 2015;7(8):2321–32.
41. Nishikawa H, Iijima T, Kajitani R, Yamaguchi J, Ando T, Suzuki Y, et al. A genetic mechanism for female-limited Batesian mimicry in *Papilio* butterfly. *Nat Genet*. 2015;47(4):405.
42. Li X, Fan D, Zhang W, Liu G, Zhang L, Zhao L, et al. Outbred genome sequencing and CRISPR/Cas9 gene editing in butterflies. *Nat Commun*. 2015;6:8212.
43. Kanost MR, Arrese EL, Cao X, Chen Y-R, Chellapilla S, Goldsmith MR, et al. Multifaceted biological insights from a draft genome sequence of the tobacco hornworm moth, *Manduca sexta*. *Insect Biochem Mol Biol*. 2016;76:118–47.
44. Shen J, Cong Q, Kinch LN, Borek D, Otwinowski Z, Grishin NV. Complete genome of *Pieris rapae*, a resilient alien, a cabbage pest, and a source of anti-cancer proteins. *F1000Res*. 2016;5:2631.
45. Pearce SL, Clarke DF, East PD, Elfekih S, Gordon K, Jermini LS, et al. Genomic innovations, transcriptional plasticity and gene loss underlying the evolution and divergence of two highly polyphagous and invasive *Helicoverpa* pest species. *BMC Biol*. 2017;15(1):63.
46. Cheng T, Wu J, Wu Y, Chilukuri RV, Huang L, Yamamoto K, et al. Genomic adaptation to polyphagy and insecticides in a major East Asian noctuid pest. *Nat Ecol Evol*. 2017;1(11):1747.
47. Wu N, Zhang S, Li X, Cao Y, Liu X, Wang Q, et al. Fall webworm genomes yield insights into rapid adaptation of invasive species. *Nat Ecol Evol*. 2019;3(1):105–15.
48. Zhang L, Liu B, Zheng W, Liu C, Zhang D, Zhao S, et al. High-depth resequencing reveals hybrid population and insecticide resistance characteristics of fall armyworm (*Spodoptera frugiperda*) invading China. *bioRxiv*. 2019;813154. <https://doi.org/10.1101/813154>.
49. Lu S, Yang J, Dai X, Xie F, He J, Dong Z, et al. Chromosomal-level reference genome of Chinese peacock butterfly (*Papilio bianor*) based on third-generation DNA sequencing and Hi-C analysis. *GigaScience*. 2019;8(11):giz128.
50. Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, et al. Phylogenomics resolves the timing and pattern of insect evolution. *Science*. 2014;346(6210):763–7.
51. Vogt RG, Große-Wilde E, Zhou J-J. The Lepidoptera odorant binding protein gene family: gene gain and loss within the GOBP/PBP complex of moths and butterflies. *Insect Biochem Mol Biol*. 2015;62:142–53.
52. Kristensen NP, Scoble MJ, Karsholt O. Lepidoptera phylogeny and systematics: the state of inventing moth and butterfly diversity. *Mol Phylogenet Evol*. 2009;43(57):237–44.
53. Mutanen M, Wahlberg N, Kaila L. Comprehensive gene and taxon coverage elucidates radiation patterns in moths and butterflies. *Proc Biol Sci*. 2010;277(1695):2839–48.
54. Kawahara AY, Plotkin D, Espeland M, Meusemann K, Toussaint EF, Donath A, et al. Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. *Proc Natl Acad Sci*. 2019;116(45):22657–63.
55. Haines T, Horley D. Walking with beasts: a prehistoric safari. *DK Pub*; 2001.
56. Zhang L-W, Kang K, Jiang S-C, Zhang Y-N, Wang T-T, Zhang J, et al. Analysis of the antennal transcriptome and insights into olfactory genes in *Hyphantria cunea* (Drury). *PLoS One*. 2016;11(10):e0164729.
57. Faye I, Pye A, Rasmuson T, Boman HG, Boman I. Insect immunity. 11. Simultaneous induction of antibacterial activity and selection synthesis of some hemolymph proteins in diapausing pupae of *Hyalophora cecropia* and *Samia cynthia*. *Infect Immun*. 1975;12(6):1426–38.
58. Steiner H, Hultmark D, Engström Å, Bennich H, Boman H. Sequence and specificity of two antibacterial proteins involved in insect immunity. *Nature*. 1981;292(5820):246.
59. Gorman MJ, Paskewitz SM. Serine proteases as mediators of mosquito immune responses. *Insect Biochem Mol Biol*. 2001;31(3):257–62.
60. WIESNER A, LOSEN S, KOPÁČEK P, WEISE C, GÖTZ P. Isolated apolipoprotein III from *Galleria mellonella* stimulates the immune reactions of this insect. *J Insect Physiol*. 1997;43(4):383–91.
61. Kamita S, Maeda S. Inhibition of *Bombyx mori* nuclear polyhedrosis virus (NPV) replication by the putative DNA helicase gene of *Autographa californica* NPV. *J Virol*. 1993;67(10):6239–45.
62. Wu Q, Brown MR. Signaling and function of insulin-like peptides in insects. *Annu Rev Entomol*. 2006;51:1–24.
63. Riakhel A, Dhadialla T. Accumulation of yolk proteins in insects oocytes. *Annu Rev Entomol*. 1992;37:217–51.
64. Izumi S, Yano K, Yamamoto Y, Takahashi SY. Yolk proteins from insect eggs: structure, biosynthesis and programmed degradation during embryogenesis. *J Insect Physiol*. 1994;40(9):735–46.
65. Boman HG, Faye I, Gudmundsson GH, Lee JY, Lidholm DA. Cell-free immunity in *Cecropia*. A model system for antibacterial proteins. *FEBS J*. 2010;201(1):23–31.
66. Dolezelova E, Zurovec M, Dolezal T, Simek P, Bryant PJ. The emerging role of adenosine deaminases in insects. *Insect Biochem Mol Biol*. 2005;35(5):381–9.
67. Edgar BA. How flies get their size: genetics meets physiology. *Nat Rev Genet*. 2006;7(12):907.
68. Sappington TW, Raikhel AS. Molecular characteristics of insect vitellogenins and vitellogenin receptors. *Insect Biochem Mol Biol*. 1998;28(5–6):277–300.
69. Raubenheimer D, Simpson SJ. Integrative models of nutrient balancing: application to insects and vertebrates. *Nutr Res Rev*. 1997;10(1):151–79.
70. Raubenheimer D, Simpson SJ. Nutrient balancing in grasshoppers: behavioural and physiological correlates of dietary breadth. *J Exp Biol*. 2003;206(10):1669–81.
71. Feyerisen R, Koener JF, Cariño FA, Daggett AS. *Biochemistry and Molecular Biology of Insect Cytochrome P450*. US: Springer; 1990. p. 263–72.
72. Zhang L, Lu Y, Xiang M, Shang Q, Gao X. The retardant effect of 2-Tridecanone, mediated by cytochrome P450, on the development of cotton bollworm, *Helicoverpa armigera*. *BMC Genomics*. 2016;17(1):954.
73. Niwa R, Matsuda T, Yoshiyama T, Namiki T, Mita K, Fujimoto Y, et al. CYP306A1, a cytochrome P450 enzyme, is essential for ecdysteroid biosynthesis in the prothoracic glands of *Bombyx* and *Drosophila*. *J Biol Chem*. 2004;279(34):35942–9.
74. Zhou H, Chen K, Yao Q, Gao L, Wang Y. Molecular cloning of *Bombyx mori* cytochrome P450 gene and its involvement in fluoride resistance. *J Hazard Mater*. 2008;160(2–3):330–6.
75. Nirmala X, Kodrik D, Zurovec M, Sehnal F. Insect silk contains both a Kunitz-type and a unique Kazal-type proteinase inhibitor. *FEBS J* 2010;268(7):2064–2073.
76. Altman GH, Diaz F, Jakuba C, Calabro T, Horan RL, Chen J, et al. Silk-based biomaterials. *Biomaterials*. 2003;24(3):401–16.
77. Zhao P, Dong Z, Duan J, Wang G, Wang L, Li Y, et al. Genome-wide identification and immune response analysis of serine protease inhibitor genes in the silkworm, *Bombyx mori*. *Plos One*. 2012;7(2):e31168.
78. Yi Q, Zhao P, Wang X, Zou Y, Zhong X, Wang C, et al. Shotgun proteomic analysis of the *Bombyx mori* anterior silk gland: an insight into the biosynthetic fiber spinning process. *Proteomics*. 2013;13(17):2657–63.
79. Inoue S, Tanaka K, Arisaka F, Kimura S, Ohtomo K, Mizuno S. Silk fibroin of *Bombyx mori* is secreted, assembling a high molecular mass elementary unit consisting of H-chain, L-chain, and P25, with a 6:6:1 molar ratio. *J Biol Chem*. 2000;275(51):40517–28.
80. Sutherland TD, Young JH, Weisman S, Hayashi CY, Merritt DJ. Insect silk: one name, many materials. *Annu Rev Entomol*. 2010;55(1):171.
81. Song F, Zhang P, Yi F, Hong X, Lu C, Yutaka B, et al. Study on fibroin heavy chain of the silkworm *Bombyx mori* by fluorescence in situ hybridization (FISH). *Sci China*. 2002;45(6):663–8.
82. Wang X, Fang X, Yang P, Jiang X, Jiang F, Zhao D, et al. The locust genome provides insight into swarm formation and long-distance flight. *Nat Commun*. 2014;5(5):2957.
83. Miller KG, Browning JV, Aubry MP, Wade BS, Katz ME, Kulpecz AA, et al. Eocene-Oligocene global climate and sea-level changes: St. Stephens quarry, Alabama. *Geol Soc Am Bull*. 2008;120(1):34–53.
84. Wolfe JA. A Paleobotanical interpretation of tertiary climates in the northern hemisphere: data from fossil plants make it possible to reconstruct tertiary climatic changes, which may be correlated with changes in the inclination of the earth's rotational axis. *Am Sci*. 1978;66(6):694–703.
85. Berggren WA, Prothero DR. *Eocene-Oligocene climatic and biotic evolution*. Princeton: Princeton University Press; 1992.

86. Mei T, Fu W-B, Li B, He Z-B, Chen B. Comparative genomics of chemosensory protein genes (CSPs) in twenty-two mosquito species (Diptera: Culicidae): identification, characterization, and evolution. *PLoS One*. 2018;13(1):e0190412.
87. Xu W, Alexie P, Zhang HJ, Alisha A. Expansion of a bitter taste receptor family in a polyphagous insect herbivore. *Sci Rep*. 2016;6:23666.
88. Li X, Schuler MA, Berenbaum MR. Molecular mechanisms of metabolic resistance to synthetic and natural xenobiotics. *Annu Rev Entomol*. 2007; 52(1):231.
89. Tsubota T, Shiotsuki T. Genomic and phylogenetic analysis of insect carboxyl/cholinesterase genes. *J Pestic Sci*. 2010;35(2):310–4.
90. Ahn S, Vogel H, Heckel D. Comparative analysis of the UDP-glycosyltransferase multigene family in insects. *Insect Biochem Mol Biol*. 2012;42(2):133–47.
91. Zhou D, Liu X, Sun Y, Ma L, Shen B, Zhu C. Genomic analysis of detoxification supergene families in the mosquito *Anopheles sinensis*. *PLoS One*. 2015;10(11):e0143387.
92. Claudianos C, Ranson H, Johnson R, Biswas S, Schuler M, Berenbaum M, et al. A deficit of detoxification enzymes: pesticide sensitivity and environmental response in the honeybee. *Insect Mol Biol*. 2006;15(5): 615–36.
93. Duman JG, Xu L, Neven LG, Tursman D, Wu DW. Hemolymph proteins involved in insect subzero-temperature tolerance: ice nucleators and antifreeze proteins. *Insects at low temperature*: Springer; 1991. p. 94–127.
94. Andreeva-Kovalevskaya ZI, Solonin A, Sineva E, Ternovsky V. Pore-forming proteins and adaptation of living organisms to environmental conditions. *Biochem Mosc*. 2008;73(13):1473–92.
95. Srinivasan A, Giri AP, Gupta VS. Structural and functional diversities in lepidopteran serine proteases. *Cell Mol Biol Lett*. 2006;11(1):132.
96. Wang J, Li DZ, Min SF, Mi F, Zhou SS, Wang MQ. Analysis of chemosensory gene families in the beetle *Monochamus alternatus* and its parasitoid *Dastarcus helophoroides*. *Comp Biochem Physiol Part D Genomics Proteomics*. 2014;11(9):1–8.
97. do ARB N, Fresia P, Cónsoli FL, Omoto C. Comparative transcriptome analysis of lufenuron-resistant and susceptible strains of *Spodoptera frugiperda* (Lepidoptera: Noctuidae). *BMC Genomics*. 2015;16(1):985.
98. Yu QY, Fang SM, Zhang Z, Jiggins CD. The transcriptome response of *Heliconius melpomene* larvae to a novel host plant. *Mol Ecol*. 2016;25(19): 4850–65.
99. Dermauw W, Wybouw N, Rombauts S, Menten B, Vontas J, Grbić M, et al. A link between host plant adaptation and pesticide resistance in the polyphagous spider mite *Tetranychus urticae*. *Proc Natl Acad Sci*. 2013; 110(2):E113–E22.
100. Strode C, Steen K, Ortelli F, Ranson H. Differential expression of the detoxification genes in the different life stages of the malaria vector *Anopheles gambiae*. *Insect Mol Biol*. 2006;15(4):523–30.
101. Rey D, Cuany A, Pautou M-P, Meyran J-C. Differential sensitivity of mosquito taxa to vegetable tannins. *J Chem Ecol*. 1999;25(3):537–48.
102. Hennigesjanssen K, Reineke A, Heckel DG, Groot AT. Complex inheritance of larval adaptation in *Plutella xylostella* to a novel host plant. *Heredity*. 2011; 107(5):421.
103. JJMe X. Invited review: microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances. *Mol Ecol*. 2006;15(7): 1713–31.
104. Shi W, Syrenne R, Sun JZ, JSJIS Y. Molecular approaches to study the insect gut symbiotic microbiota at the 'omics' age. *Insect Science*. 2010; 17(3):199–219.
105. Colman DR, Toolson EC, CJME T-V. Do diet and taxonomy influence insect gut bacterial communities? *Mol Ecol*. 2012;21(20):5124–37.
106. Xiang H, Wei G-F, Jia S, Huang J, Miao X-X, Zhou Z, et al. Microbial communities in the larval midgut of laboratory and field populations of cotton bollworm (*Helicoverpa armigera*). *Can J Microbiol*. 2006;52(11):1085–92.
107. Xiang H, Li M, Zhao Y, Zhao L, Zhang Y, Huang Y. Bacterial community in midguts of the silkworm larvae estimated by PCR/DGGE and 16S rDNA gene library analysis; 2007.
108. Delalibera I Jr, Handelsman J, KFJEE R. Contrasts in cellulolytic activities of gut microorganisms between the wood borer, *Saperda vestita* (Coleoptera: Cerambycidae), and the bark beetles, *Ips pini* and *Dendroctonus frontalis* (Coleoptera: Curculionidae). *Environ Entomol*. 2005;34(3):541–7.
109. Dixon R, DJNRM K. Genetic regulation of biological nitrogen fixation. *Nat Rev Microbiol*. 2004;2(8):621.
110. Behar A, Yuval B, EJME J. Enterobacteria-mediated nitrogen fixation in natural populations of the fruit fly *Ceratitis capitata*. *Mol Ecol*. 2005;14(9): 2637–43.
111. Delalibera I Jr, Handelsman J, Raffa KF. Contrasts in cellulolytic activities of gut microorganisms between the wood borer, *Saperda vestita* (Coleoptera: Cerambycidae), and the bark beetles, *Ips pini* and *Dendroctonus frontalis* (Coleoptera: Curculionidae). *Environ Entomol*. 2005;34(3):541–7.
112. Dixon R, Kahn D. Genetic regulation of biological nitrogen fixation. *Nat Rev Microbiol*. 2004;2(8):621.
113. Behar A, Yuval B, Jurkevitch E. Enterobacteria-mediated nitrogen fixation in natural populations of the fruit fly *Ceratitis capitata*. *Mol Ecol*. 2005;14(9): 2637–43.
114. Dillon RJ, Dillon VM. The gut bacteria of insects: nonpathogenic interactions. *Annu Rev Entomol*. 2004;49(1):71–92 PubMed PMID: 14651457. Epub 2003/ 12/04.
115. Chen B, Yu T, Xie S, Du K, Liang X, Lan Y, et al. Comparative shotgun metagenomic data of the silkworm *Bombyx mori* gut microbiome. *Sci Data*. 2018;5:180285.
116. Suen G, Scott JJ, Aylward FO, Adams SM, Tringe SG, Pinto-Tomas AA, et al. An insect herbivore microbiome with high plant biomass-degrading capacity. *PLoS Genet*. 2010;6(9):e1001129 PubMed PMID: 20885794. Pubmed Central PMCID: PMC2944797. Epub 2010/10/05.
117. Pinto-Tomas AA, Anderson MA, Suen G, Stevenson DM, FST C, Cleland WW, et al. Symbiotic Nitrogen Fixation in the Fungus Gardens of Leaf-Cutter Ants. *Science*. 2009;326(5956):1120–3 PubMed PMID: WOS: 000271951000047. English.
118. Yadav AN, Sharma D, Gulati S, Singh S, Dey R, Pal KK, et al. Haloarchaea endowed with phosphorus solubilization attribute implicated in phosphorus cycle. *Sci Rep*. 2015;5:12293.
119. Yadav AN, Verma P, Kaushik R, Dhaliwal H, AJEM S. Archaea endowed with plant growth promoting attributes. *EC Microbiol*. 2017;8(6):294–8.
120. Kikuchi Y, Hosokawa T, TJA F. Insect-microbe mutualism without vertical transmission: a stinkbug acquires a beneficial gut symbiont from the environment every generation. *Appl Environ Microbiol*. 2007;73(13):4308–16.
121. Kikuchi Y, Hayatsu M, Hosokawa T, Nagayama A, Tago K, Fukatsu T. Symbiont-mediated insecticide resistance. *Proc Natl Acad Sci U S A*. 2012; 109(22):8618–22.
122. Keeling CI, JJNP B. Genes, enzymes and chemicals of terpenoid diversity in the constitutive and induced defence of conifers against insects and pathogens. *New Phytol*. 2006;170(4):657–75.
123. Minard G, Mavingui P, Moro CV. Diversity and function of bacterial microbiota in the mosquito holobiont. *Parasit Vectors*. 2013;6(1):146.
124. Xia X, Gurr GM, Vasseur L, Zheng D, Zhong H, Qin B, et al. Metagenomic sequencing of diamondback moth gut microbiome unveils key holobiont adaptations for herbivory. *Front Microbiol*. 2017;8:663.
125. Sticklen MB. Plant genetic engineering for biofuel production: towards affordable cellulosic ethanol. *Nat Rev Genet*. 2008;9(6):433.
126. Liu N, Zhang L, Zhou H, Zhang M, Yan X, Wang Q, et al. Metagenomic insights into metabolic capacities of the gut microbiota in a fungus-cultivating termite (*Odontotermes yunnanensis*). *PLoS One*. 2013;8(7): e69184.
127. Patel DD, Patel AK, Parmar NR, Shah TM, Patel JB, Pandya PR, et al. Microbial and Carbohydrate Active Enzyme profile of buffalo rumen metagenome and their alteration in response to variation in the diet. *Gene*. 2014;545(1):88–94.
128. USA MDoC. Fall webworm *Hyphantria cunea* (Drury). External Factsheets. 2000.
129. Fitzgerald TD. Sociality in caterpillar; 1993.
130. Takuya T, Kimiko Y, Kazuei M, et al. Gene expression analysis in the larval silk gland of the eri silkworm *Samia ricini*. *Insect Sci*. 2016;23(6):791–804.
131. Rehnberg BG. Temperature profiles inside webs of the fall webworm, *Hyphantria cunea* (Lepidoptera: Arctiidae): Influence of weather, compass orientation, and time of day. *J Therm Biol*. 2006;31(3):274–9.
132. Mori H, Tsukada M. New silk protein: modification of silk protein by gene engineering for production of biomaterials. *Rev Mol Biotechnol*. 2000;74(2): 95–103.
133. Kato N, Sato S, Yamanaka A, Yamada H, Fuwa N, Nomura M. Silk protein, sericin, inhibits lipid peroxidation and tyrosinase activity. *J Agr Chem Soc Jpn*. 1998;62(1):145–7.
134. Terada S, Nishimura T, Sasaki M, Yamada H, Miki M. Sericin, a protein derived from silkworms, accelerates the proliferation of several mammalian cell lines including a hybridoma. *Cytotechnology*. 2002;40(1–3):3–12.

135. Terada S, Sasaki M, Yanagihara K, Yamada H. Preparation of silk protein sericin as mitogenic factor for better mammalian cell culture. *J Biosci Bioeng.* 2005;100(6):667–71.
136. Morikawa M, Kimura T, Murakami M, Katayama K, Terada S, Yamaguchi A. Rat islet culture in serum-free medium containing silk protein sericin. *J Hepatobiliary Pancreatic Surg.* 2009;16(2):223–8.
137. Manosroi A, Boonpisuttinant K, Winitchai S, Manosroi W, Manosroi J. Free radical scavenging and tyrosinase inhibition activity of oils and sericin extracted from Thai native silkworms (*Bombyx mori*). *Pharm Biol.* 2010;48(8): 855–60.
138. Yang M. Silk-based biomaterials. *Microsc Res Tech.* 2017;80(3):321–30.
139. Li S, Liu B, Cheng J, Hu J. Composite cement of magnesium-bearing phosphoaluminate-hydroxyapatite reinforced by treated raw silk fiber. *Cement Concrete Composites.* 2008;30(4):347–52.
140. Oyama F, Mizuno S, Shimura K. Studies on immunological properties of fibroin heavy and light chains. *J Biochem.* 1984;96(6):1689–94.
141. Matsuno K, Hui CC, Takiya S, Suzuki T, Ueno K, Suzuki Y. Transcription signals and protein binding sites for sericin gene transcription in vitro. *J Biol Chem* 1989;264(31):18707–18713.
142. Matsuno K, Takiya S, Hui CC, Suzuki T, Fukuta M, Ueno K, et al. Transcriptional stimulation via 5C site of *Bombyx sericin-1* gene through an interaction with a DNA binding protein SGF-3. *Nucleic Acids Res.* 1990;18(7): 1853–8.
143. Ohno K, Sawada JI, Takiya S, Mai K, Matsumoto A, Tsubota T, et al. Silk Gland Factor-2 (SGF-2) Involved in Fibroin Gene Transcription Consists of LIM-homeodomain, LIM-interacting, and Single-Stranded DNA-Binding Proteins. *J Biol Chem.* 2013;288(44):31581.
144. Tsuda M, Suzuki Y. Faithful transcription initiation of fibroin gene in a homologous cell-free system reveals an enhancing effect of 5' flanking sequence far upstream. *Cell* 1981;27(1):175–182.
145. Takiya S, Hui CC, Suzuki Y. A contribution of the core-promoter and its surrounding regions to the preferential transcription of the fibroin gene in posterior silk gland extracts. *Embo J.* 1990;9(2):489–496.
146. Guo PC, Dong Z, Zhao P, Zhang Y, He H, Tan X, et al. Structural insights into the unique inhibitory mechanism of the silkworm protease inhibitor serpin18. *Sci Rep.* 2015;5:11863.
147. Zhang Y, Zhao P, Dong Z, Wang D, Guo P, Guo X, et al. Comparative proteome analysis of multi-layer cocoon of the silkworm, *Bombyx mori*. *Plos One.* 2015;10(4):e0123403.
148. Biosciences P. Procedure & Checklist—20 kb Template Preparation Using BluePippinTM Size Selection System. SampleNet; 2014.
149. Hale CM, Chen W-C, Khatau SB, Daniels BR, Lee JS, Wirtz D. SMRT analysis of MTOC and nuclear positioning reveals the role of EB1 and LIC1 in single-cell polarization. *J Cell Sci.* 2011;124(Pt 24):4267–85.
150. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 2017;27(5):722–36.
151. Shigang W. A fuzzy Bruijn graph approach to long noisy reads assembly 2017 [cited 2018 12th,oct]. Available from: <https://github.com/ruanjue/wtdbg>.
152. De Landtsheer S, Trairatphisan P, Lucarelli P, TJB S. FALCON: a toolbox for the fast contextualization of logical networks. *Bioinformatics.* 2017;33(21):3431–6.
153. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One.* 2014;9(11):e112963 PubMed PMID: 25409509. PubMed Central PMCID: PMC4237348. Epub 2014/11/20.
154. Parra G, Bradnam K, IJB K. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics.* 2007;23(9):1061–7.
155. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM, et al. *Bioinformatics.* 2015;31(19):3210–2.
156. Xu Z, HJNar W. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 2007;35(suppl_2):W265–W8.
157. Han Y, Wessler SR. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* 2010;38(22):e199.
158. Price AL, Jones NC, Pevzner PAJB. De novo identification of repeat families in large genomes. *Bioinformatics.* 2005;21(suppl_1):i351–i8.
159. Edgar RC, Myers EWJB. PILER: identification and classification of genomic repeats. *Bioinformatics.* 2005;21(suppl_1):i152–i8.
160. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet.* 2007;8(12):973.
161. Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics.* 2004;5(1):4.10 1–4. 4.
162. Lowe TM. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Eddy SR.* 1997;25(5):955.
163. Nawrocki EP, Eddy SRJB. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics.* 2013;29(22):2933–5.
164. She R, Chu JS-C, Wang K, Pei J, Chen N. GenBlastA: enabling BLAST to identify homologous gene sequences. *Genome Res.* 2009;19(1):143–9.
165. Birney E, Clamp M, Durbin R. GeneWise and genomewise. *Genome Res.* 2004;14(5):988–95.
166. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA1. *J Mol Biol.* 1997;268(1):78–94.
167. Stanke M, Waack SJB. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics.* 2003;19(suppl_2):ii215–i25.
168. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics.* 2004;20(16):2878–9.
169. Blanco E, Parra G, Guigó R. Using geneid to identify genes. *Curr Protoc Bioinformatics.* 2007;18(1):4.3 1–4.3. 28.
170. Korf I. Gene finding in novel genomes. *BMC Bioinformatics.* 2004;5(1):59.
171. Keilwagen J, Wenk M, Erickson JL, Schattat MH, Grau J, Hartung F. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res.* 2016;44(9):e89.
172. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc.* 2016;11(9):1650.
173. Haas B, Papanicolaou AJGS. TransDecoder (find coding regions within transcripts); 2016.
174. Tang S, Lomsadze A, Borodovsky M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res.* 2015;43(12):e78.
175. Campbell MA, Haas BJ, Hamilton JP, Mount SM, Buell CR. Comprehensive analysis of alternative splicing in rice and comparative analyses with *Arabidopsis*. *BMC Genomics.* 2006;7(1):327.
176. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EvidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* 2008;9(1):1.
177. Chen F, Mackey AJ, Stoeckert CJ Jr, Roos DS. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* 2006;34(suppl_1):D363–D8.
178. Löytynoja A, Goldman N. An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci U S A.* 2005; 102(30):10557–62.
179. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 2000;17(4):540–52.
180. Shao Y, Li J-X, Ge R-L, Zhong L, Irwin DM, Murphy RW, et al. Genetic adaptations of the plateau zokor in high-elevation burrows. *Sci Rep.* 2015;5: 17262.
181. Stamatakis AJB. RAXML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics.* 2014;30(9):1312–3.
182. Sanderson MJJB. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics.* 2003; 19(2):301–2.
183. Durden CJ, Rose H. Butterflies from the middle Eocene: the earliest occurrence of fossil Papilionoidea (Lepidoptera). Texas Memorial Museum, The University of Texas at Austin; 1978.
184. Lukashevich ED, Przhiboro AA, Marchal-Papier F, Grauvogel-Stamm L. The oldest occurrence of immature Diptera (Insecta). Middle Triassic, France: Annales de la Société entomologique de France, Taylor & Francis Group. 2010;46(1-2):4-22.
185. Kirejtshuk AG, Poschmann M, Prokop J, Garroute R, Nel A. Evolution of the elytral venation and structural adaptations in the oldest Palaeozoic beetles (Insecta: Coleoptera: Tsherkardocoleidae). *J Syst Palaeontology.* 2014;12(5): 575–600.
186. De Bie T, Cristianini N, Demuth JP, Hahn MW. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics.* 2006;22(10):1269–71.
187. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 2007;24(8):1586–91.
188. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods.* 2015;12(4):357.
189. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated

- transcripts and isoform switching during cell differentiation. *Nature Biotechnol.* 2010;28(5):511.
190. Robinson MD, DJ MC, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* 2010;26(1):139–40.
 191. Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BM, et al. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics.* 2013;29(8):1035–43.
 192. Haynes W. Benjamini–hochberg method. *Encyclopedia of systems biology*; 2013. p. 78.
 193. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 2014;15(3):R46.
 194. Ondov BD, Bergman NH, Phillippy AM. Genomes, Metagenomes: Basics M, Databases, Tools. Krona: Interactive Metagenomic Visualization in a Web Browser; 2015. p. 339–46.
 195. Peng Y, Leung HC, Yiu S-M, Chin FY. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics.* 2012;28(11):1420–8.
 196. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUASt: quality assessment tool for genome assemblies. *Bioinformatics.* 2013;29(8):1072–5.
 197. Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* 2010;38(12):e132.
 198. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics.* 2012;28(23):3150–2.
 199. Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, et al. HMMER web server: 2015 update. *Nucleic Acids Res.* 2015;43(W1):W30–W8.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

