

Article

# Reference genome sequences of the oriental armyworm, *Mythimna separata* (Lepidoptera: Noctuidae)

Kakeru Yokoi <sup>1,\*</sup>, Seiichi Furukawa <sup>2</sup>, Rui Zhou <sup>3</sup>, Akiya Jouraku <sup>1</sup> and Hidemasa Bono <sup>4,5</sup>

<sup>1</sup> Insect Design Technology Group, Division of Insect Advanced Technology, Institute of Agrobiological Sciences, National Agriculture and Food Research Organization (NARO), 1-2 Owashi, Tsukuba, Ibaraki, Japan; joraku@affrc.go.jp

<sup>2</sup> Institute of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Ibaraki, Japan; furukawa.seiichi.ew@u.tsukuba.ac.jp

<sup>3</sup> Degree Program in Agro-Bioresources Science and Technology, University of Tsukuba, Tsukuba, Ibaraki, Japan; syuuruizz@gmail.com

<sup>4</sup> Laboratory of Genome Informatics, Graduate School of Integrated Sciences for Life, Hiroshima University, 3-10-23 Kagamiyama, Higashi-Hiroshima City, Hiroshima 739-0046, Japan; bonohu@hiroshima-u.ac.jp

<sup>5</sup> Laboratory of BioDX, Genome Editing Innovation Center, Hiroshima University, 3-10-23 Kagamiyama, Higashi-Hiroshima City, Hiroshima, Japan;

\* Correspondence: yokoi123@affrc.go.jp; Tel.: +81-29-838-6129

**Simple Summary:** The oriental armyworm, *Mythimna separata*, an agricultural pest, is commonly used to study insect immune reactions and interactions with parasitoid wasps. To promote such studies, a reference genome was constructed. The *M. separata* genome is 682Mbp long—a size comparable to that of other lepidopteran insects. The contig N50 value of the genome is 2.7 Mb, which indicates sufficient quality to be used as reference genome data. Gene set data were constructed using genome and RNA-sequencing data; a total of 21,970 genes and 24,452 coding sites were predicted. Functional gene annotation was performed using the predicted amino acid sequences and reference gene set data of the model organism and other insect species as well as Unigene and Pfam datasets. Consequently, 45–80% of the amino acid sequences were annotated using these data sets. Using these data, most of the orthologs of core components in the Toll and immune deficiency (IMD) pathways were identified, suggesting the presence of these two pathways in *M. separata*. Additionally, 105 C-type lectins were identified in the *M. separata* genome, which were more numerous than those in other insect species, suggesting that these genes may be duplicated.

**Abstract:** Lepidopteran insects are an important group of animals, among which some are used as biochemical and physiological model species in the insect and silk industries, whereas others are major agricultural pests. Therefore, genome sequences of several lepidopteran insects have been reported thus far. The oriental armyworm, *Mythimna separata*, is an agricultural pest commonly used to study insect immune reactions and interactions with parasitoid wasps as hosts. To improve our understanding of these research topics, reference genome sequences were constructed in the present study. Using long-read and short-read sequence data, *de novo* assembly and polishing were performed, and haplotigs were purged. Subsequently, gene predictions and functional annotations were performed. To search for orthologs of the Toll and immune deficiency (IMD) pathways and C-type lectins, annotation data analysis, BLASTp, and Hummer scans were performed. The *M. separata* genome is 682 Mbp; its contig N50 was 2.7 Mbp with 21,970 genes and 24,452 coding sites predicted. All orthologs of the core components of the Toll and IMD pathways and 105 C-type lectins were identified. These results suggest that the genome data were of sufficient quality as reference genome data and could contribute to promoting *M. separata* and lepidopteran research at the molecular and genome levels.

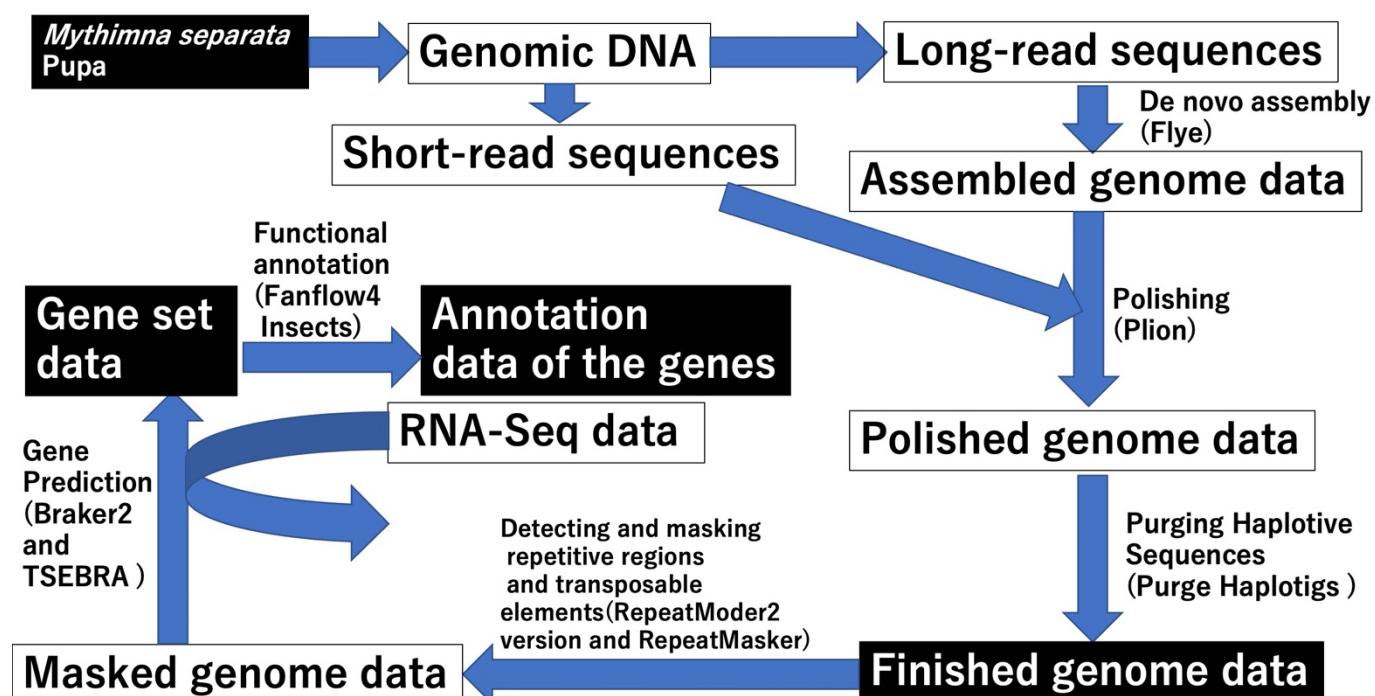
**Keywords:** Genome sequence; *Mythimna separata*; Toll pathway; IMD pathway; C-type lectin

## 1. Introduction

Several species of lepidopteran insects are known, of which some are beneficial and some detrimental. Although the domestic silkworm (*Bombyx mori*; Bombycidae; Linnaeus, 1758), is used for silk production, some lepidopteran insects are major agricultural pests. For example, the fall armyworm (*Spodoptera frugiperda*; Noctuidae; J. E. Smith, 1797), the beet armyworm (*Spodoptera exigua*; Noctuidae; Hübner, 1808), the tobacco cutworm (*Spodoptera litura*; Noctuidae; Fabricius, 1775), The diamondback moth (*Plutella xylostella*; Plutellidae; Linnaeus, 1758), the cabbage looper (*Trichoplusia ni*; Noctuidae; Hübner, 1800–1803), and the oriental armyworm (*Mythimna separata*; Noctuidae; Walker, 1865) are the most severe pests in some crops [1–3]. In contrast, the tobacco hornworm (*Manduca sexta*; Sphingidae; Linnaeus, 1763) has been used as a biochemical and physiological model species of insects (e.g., to study immune reaction, development, and metamorphosis [4,5]). Because of this importance, research on the lepidopteran family of insects is quite active, and the genomes of some lepidoptera species have been sequenced. The first lepidopteran genome sequence data to be reported was the draft genome sequence data of *Bombyx mori* in 2004 [6,7]. The genome data was updated [8] and, the chromosome-level genome sequence of *B. mori* was reported [9]. Additionally, genome sequence data of other lepidopteran families, including *S. frugiperda* [10,11], *S. exigua* [12], *S. litura* [2], *P. xylostella* [13,14], *T. ni* [15], and *M. sexta*, have been reported [16,17]. In addition, the genome sequence data for several butterflies were reported [18].

*M. separata* is commonly used to study immune reactions in insects. Insects exhibit humoral and cellular immune responses to foreign microorganisms [19]. *M. separata* larval hemocytes, especially granular cells and plasmacytocytes, play a central role in cellular immune reactions, and different types of invaders activate different reactions such as encapsulation, phagocytosis, and nodule formation. Several molecules in the *M. separata* hemolymph mediate these reactions. Growth blocking peptides, found as insect cytokines, trigger the spreading of plasmacytocytes [20]. C-type lectins (CTLs) are a superfamily of proteins that recognize carbohydrates in a calcium-dependent manner. Some CTLs can activate cellular immune reactions. Ishihara et al. (2017) reported that the CTL encapsulation promoting lectin (EPL) enhances encapsulation [21]. Another CTL, IML-10, also promotes encapsulation [22]. A transcriptome analysis found 35 CTLs from *M. separata* larvae [23], and many of them changed their expression profiles upon bacterial exposure, suggesting that they are involved in immune regulation. To study the interactions between *M. separata* and parasitic wasps or flies, the oriental armyworm is used as a host insect for these species [24–28]. Using both species, our group identified several candidate factors needed for successful parasites in the braconid wasp, *Meteorus pulchricornis* (Hymenoptera: Braconidae; Haliday, 1835) [29] and investigated several gene functions of *M. separata* related to apoptosis or immune reactions considered to be essential for parasitoids [30,31].

To deepen our understanding of the molecular mechanisms of immune reactions and interactions between *M. separata* and parasitoid wasps, we constructed a reference-quality genome sequence of the oriental armyworm (*M. separata*), using long-read and short-read sequence data. Consequently, gene set data and the translated amino acid sequence data were prepared, and functional annotations of the predicted genes were performed (Figure 1). Using the gene set data, the orthologues consisting of IMD and Toll pathways, which are major immune signaling pathways, were searched. Finally, CTL genes were searched in the gene sets, and sequence and domain analysis of CTL were performed. These results suggested that the reference genome data and gene set data could contribute to promoting the *M. separata* and Lepidopteran research at the molecular or genome level.



**Figure 1.** Schemes of the reference genome constructions, gene prediction and functional annotating. “Finished genome data” indicates the reference genome data. The names of software used in each step are shown in brackets

94  
95  
96  
97  
98  
99

## 2. Materials and Methods

### 2.1 Sample preparation and sequencing

*M. separata* was supplied from stock cultures stored at Takeda Chemical Industries, Ltd. [32] and maintained in the Laboratory of Applied Entomology and Zoology, University of Tsukuba, Japan. The insect was reared on an artificial diet (Silkmate, Nihon Nosen Kogyo, Kanagawa, Japan) at  $25\pm2$  °C, 40–80% relative humidity (r. h.), and an L16:D8 photoperiod.

Genomic DNA from single male pupae was extracted using NucleoBond HMW DNA (TaKaRa Bio Inc, Shiga, Japan) according to manufacturer's protocol. Briefly, the whole body in 500  $\mu$ L Lysis buffer was homogenized with a plastic pestle and liquid nitrogen. After treatments of 200  $\mu$ L Liquid Proteinase K at 50 °C for 2 h and 100  $\mu$ L Liquid RNase A at 25 °C for 5 min, the homogenate was loaded onto NucleoBond HMW Column. Through washing step, DNA was finally eluted with 5  $\mu$ L Elution buffer. Prep Kit 2.0 (Pacific Bioscience, CA, USA) according to manufacturer's protocol. The sequencing library was size-selected using the BluePippin system (Sigma Science, MA, USA) with a lower cutoff of 30 kbp. One SMRT Cell 8M was sequenced on the PacBio Sequel II System with Binding Kit 2.0 and Sequencing Kit 2.0, yielding a total of 6,971,329 polymerase reads (201,369,397,241 bp). Shearing System M220 (Covaris Inc., MA, USA). A paired-end library was constructed with a TruSeq DNA PCR-Free Library Prep kit (Illumina, CA, USA) and was size-selected on an agarose gel using a Zymoclean Large Fragment DNA Recovery Kit (Zymo Research, CA, USA). The final library was sequenced on the Illumina NovaSeq 6000 sequencer with a read length of 150 bp.

Total RNA samples were prepared from the hemocytes of *M. separata* larvae, using TRIzol (Thermo Fisher Scientific, Waltham, MA, USA). PolyA RNA libraries were constructed using the TruSeq Stranded mRNA Library Prep Kit (Illumina), and sequencing (paired-end sequencing with 100-nt reads) was performed on a NovaSeq6000 platform (Illumina).

All raw sequence data were deposited in the Sequence Read Archive (SRA) of the DNA Data Bank of Japan (DDBJ). SRA accession IDs of the raw sequence data used in this study are listed in Supplementary Data 1.

### 2.2 Genome assembly and gene prediction

The output BAM file of Sequel II was converted to FASTA file using BAM2fastx version 1.3.1 (URL: <https://github.com/PacificBiosciences/bam2fastx>, accessed on 2 September 2022). *De novo* assembly was performed using the converted FASTA file with Flye version 2.9-b1774 with default settings [33]. Polishing of the assembled genome was performed using Minimap2 version 2.17 (mapping) [34], SAMtools version 1.10 (file conversion) [35], and Plion version 1.23 (perform polishing) [36]. Identification of the allelic contig pairings and productions haplotype-fused assemblies were performed using Purge Haplots version 1.1.2. The genome sequences was assessed using BUSCO version5.2.2 with eukaryota\_odb10 file [37]. Status of the genome sequences were evaluated using assembly-stats version 1.0.1 (URL: <https://github.com/sanger-pathogens/assembly-stats> accessed on 2 September 2022). The finished genome sequences were deposited to DDBJ/ENA/GenBank (accession IDs: BSAF01000001- BSAF01000569; Supplemental Data 2)

To search for repetitive sequences and transposable elements (TEs), RepeatMod2 version DEV and RepeatMasker version 4.1.2-p1 were used [38,39]. Gene prediction was performed using masked sequences and RNA-sequencing (RNA-Seq) data, including those from the public database as hint data (Supplementary Data 1). RNA-Seq data mappings to the reference genome were performed using HISAT2 version 2.2.1 [40] and mapped data were converted using SAMtools. Gene predictions were performed twice using Braker2 version 2.1.6 with RNA-Seq data-hint only mode and protein data-hint only mode independently[41]. The two output data from RNA-Seq data-hint only mode and protein data-hint only mode were merged by TSEBRA version 1.0.3 [42]. The amino acid sequences of the gene sets were functionally annotated by the functional annotation workflow Fanflow4Insects [43]. The functional annotation included the assignment of the top hit genes from comprehensively annotated organisms using the sequence similarity search program in global alignment (GGSEARCH) (<https://fasta.biocb.virginia.edu/>; accessed on 1 August 2022) and the protein domains using the protein domain database Pfam (<http://pfam.xfam.org/>; accessed on 1 August 2022) via HMMSCAN in the HMMER package (<http://hmmer.org/>; accessed on 1 August 2022). Accession IDs of raw read sequences of the RNA-Seq is listed in Supplemental Data 1. To remove the contaminating microbe sequences, the predicted amino acid sequences were annotated using BLAST nr database (Downloaded via the URL: <https://ftp.ncbi.nlm.nih.gov/blast/db/> accessed at March 3, 2022, 444,154,253 protein sequences.) by BLAST search (version 2.12.0+).

### 2.3 Searching for and sequence analysis of Immune-related genes in *M. separata*

To search for genes related to IMD and Toll pathway, BLASTx or BLASTp (version 2.2.31+) search were performed using the fruit fly, *Drosophila melanogaster* (Diptera; Meigen, 1830), sequences as query sequences from FlyBase (URL: <https://flybase.org/> accessed October 1, 2022) and predicted amino acid sequence of *M. separata* as subject sequences. Phylogenetic tree construction (Neighbor join (NJ) method with jacard methods) and domain analysis of Peptidoglycan recognition proteins (PGRPs) and Gram-negative binding protein (GNBP) (hmmscan version 3.1b2 with Pfam-A.hmm of Pfam\_35) were performed by DoMosaics (URL: <https://domainworld.uni-muenster.de/developing/domosaics/> version 0.95) [44] while other phylogenetic trees of PGRPs and GNBP (Maximum Likelihood method (ML)) were constructed by MEGA X with default settings [45]. Phylogenetic tree construction (Neighbor join method) and domain analysis of CTLs were performed by DoMosaics as same settings of the analyzing PGRPs and GNBP. SignalP (version 6.0) and DeepTMHMM (version 1.0) were used to search for signal peptide sequences and transmembrane domains in CTLs, respectively [46,47].

## 3. Results

### 3.1. Construction of the genome sequences of *M. separata*

Using long-read and short-read genome sequence data of *M. separata*, we constructed *de novo* genome sequences. The genome construction scheme is shown in Figure 1. The assembled genome sequence data were first constructed using only the long-read sequence data, with a mean coverage of 187x. Subsequently, the assembled genome sequence data was polished using short-read sequence data. The basal status of the polished genome sequence data is shown in Table 1. The total length of the polished genome data is not unusual as a lepidopteran genome with a high contig N50 value (2,745,150 bp), considering the genome sizes of lepidoptera species. Subsequently, BUSCO assessed the completeness of the polished genome data (Table 2). Although most of the core genes in the BUSCO dataset were found in the polished genome data (Complete BUSCO), some core genes, which are single-copy genes, were found to be duplicated. The reason for the duplications could be the high regional heterogeneities in diploid genomes, which eventually led to incorrect assembly of two contigs that were assembled as one contig [48]. Therefore, to address this problem, the polished genome data were loaded into the Purge Haplotigs. The improved genome data from Purge Haplotigs (referred to as the finished genome data in Table 2) were analyzed using BUSCO (Table 2). Almost all core genes in the eukaryota\_odb10 or lepidoptera\_odb10 data set were found to be complete and single-copy BUSCO in the improved genome data, indicating that the genome data had good qualities as reference genome data, with a CG percentage of 38.60%. Hereafter, we refer to these genome data as the reference genome data or simply "the genome data" (deposited in DDBJ/ENA/GenBank; see Material and Methods 2.2).

**Table 1.** The status of the constructed genome sequence data\*

Genome sequence name	Polished genome data	Finished genome data (Reference genome data)
Total length (bp)	842,875,911	681,943,315
Contig number	991	569
Contig N50 (bp)	2,220,816	2,745,150
Largest contig length (bp)	12,409,004	12,409,004
Average of contig length (bp)	850,530.69	1,198,494.40

\*Both data contain no N sequence and gap.

**Table 2.** The results of BUSCO using the constructed genome sequence data

Genome sequence name (BUSCO data set)	Polished genome data (eukaryota_odb10)	Finished genome data (eukaryota_odb10)	Finished genome data (lepidoptera_odb10)
Complete BUSCOs	253 (99.2%)	252 (98.8%)	5213 (98.6%)
Complete and single-copy BUSCOs	178 (69.8%)	251 (98.4%)	5185 (98.1%)
Complete and duplicated BUSCOs	75 (29.4%)	1 (0.4%)	28 (0.5%)
Fragmented BUSCOs	1 (0.4%)	2 (0.8%)	20 (0.4%)
Missing BUSCOs	1 (0.4%)	1 (0.4%)	53 (1.0%)

Total BUSCO groups searched	255	255	5286	199
				200
				201

### 3.2. Search for Repetitive region and Gene prediction

We searched for the repetitive regions and transposable elements (TEs) in the genome sequence (Figure 1). First, *de novo* TE detection was performed, and 1,905 consensus sequences of TEs were constructed (Supplemental Data 3). TEs and repetitive regions in the *M. separata* genome were identified using consensus sequences; the summary status is shown in Table 3, and the output files are provided in Supplementary Data 4. Approximately 46.59% of the *M. separata* genome was repetitive or contained TE sites. Among the annotated TEs, the number of retroelements (Class I TEs) was much larger than that of DNA transposons (Class II TEs). Long interspersed nuclear elements (LINEs) were the most abundant retroelements, whereas Tc1-IS630-Pogo was the most abundant DNA transposon. Rolling circles covered 6.86% of the *M. separata* genome, whereas satellites and simple repeats were less than 1%.

**Table 3.** Transposons and repetitive regions in *M. separata* genome.

	total length: bases masked:	681943315 bp 317706294 bp ( 46.59 %)	Number of elements*	length occupied	Percentage of sequence
Retroelements			483,943	102,500,601 bp	15.03%
SINEs:			69	3,659 bp	0.00%
Penelope			0	0 bp	0.00%
LINEs:			466,443	94,162,645 bp	13.81%
CRE/SLACS			11,265	2,482,075 bp	0.36%
L2/CR1/Rex			47,274	15,582,975 bp	2.29%
R1/LOA/Jockey			125,283	28,136,850 bp	4.13%
R2/R4/NeSL			8,079	2,985,258 bp	0.44%
RTE/Bov-B			184,015	31,548,956 bp	4.63%
L1/CIN4			0	0 bp	0.00%
LTR elements:			17,431	8,334,297 bp	1.22%
BEL/Pao			1,963	3,457,738 bp	0.51%
Ty1/Copia			2,715	1,358,093 bp	0.20%
Gypsy/DIRS1			6,472	2,769,678 bp	0.41%
Retroviral			160	76,839 bp	0.01%
DNA transposons			80,267	20,701,313 bp	3.04%
hobo-Activator			6,121	1,085,735 bp	0.16%
Tc1-IS630-Pogo			24,800	10,838,060 bp	1.59%
En-Spm			0	0 bp	0.00%
MuDR-IS905			0	0 bp	0.00%
PiggyBac			166	236,668 bp	0.03%
Tourist/Harbinger			2,055	550,091 bp	0.08%
Other (Mirage, P-element, Transib)			110	177,033 bp	0.03%
Rolling-circles			281,101	46,800,852 bp	6.86%
Unclassified:			880,600	14,0919,871 bp	20.66%
Total interspersed repeats:				26,4121,785 bp	38.73%
Small RNA:			204	23,041 bp	0.00%
Satellites:			177	69,651 bp	0.01%
Simple repeats:			102,349	5,964,291 bp	0.87%
Low complexity:			15,392	726,674 bp	0.11%

\* most repeats fragmented by insertions or deletions have been counted as one element

Gene predictions of *M. separata* were performed using the masked genome data (Supplementary Data 1) and RNA-Seq data as hint data, of which some were retrieved from SRA public database (Supplementary Data 1 and Figure 1). A total of 21,970 genes were identified in the *M. separata* genome (gene.gtf in Supplementary Data 5), whereas 24,452 coding sequences (CDSs) and amino acid sequences were predicted because some genes had multiple transcripts (gene\_cds.fasta and gene\_pep.fasta in Supplementary Data 5). Using the predicted amino acid sequences, the predicted genes were functionally annotated with Fanflow4Insects (Supplementary Data 6)[43]. Approximately 45–57% of the protein sequences were annotated using gene sets of model species (*Homo sapiens*, *Mus musculus*, and *Caenorhabditis elegans* (Maupas, 1900)). Among the insect species datasets, a relatively higher percentage of protein sequences were annotated using Lepidoptera datasets (*M. sexta* and *B. mori*) than using other insect species datasets (*D. melanogaster*, the western honeybee (*Apis mellifera*; Hymenoptera; Linnaeus, 1758), and the red flour beetle (*Tribolium castaneum*; Coleoptera; Herbst, 1797)), and over half of the sequences were annotated using Pfam and Unigene datasets.

**Table 4.** Numbers of the hit genes in each data set by Fanflow4.

	Numbers of the hit proteins (among 24,453 proteins)	Percentage of the hit proteins (%)
<i>H. sapience</i>	14,037	57.4039995
<i>M. musculus</i>	13,699	56.021756
<i>C. elegans</i>	11,235	45.9452828
<i>D. melanogaster</i>	12,879	52.6683842
<i>B. mori</i>	19,704	80.5790701
<i>M. sexta</i>	17,996	73.594242
<i>A. mellifera</i>	14,427	58.9988958
<i>T. castaneum</i>	16,496	67.4600254
Unigene	16,086	65.7833395
Pfam	13,244	54.1610436

We checked whether the microbial sequences were contaminated in the reference genome and attempted to remove the microbial sequences using annotation data of the gene set with NCBI-nr (Supplementary Data 7). We counted the numbers of the names of the microbial species in the annotation data (Supplementary Data 7). A total of 32, 23, 39, 12, and 10 genes were annotated with orthologs from *Pseudomonas aeruginosa*, *Piscirickettsia salmonis*, Spodoptera moth adenovirus 1, *Trichoplusia ni* TED virus, and *Conidiobolus coronatus*, respectively. If these genes are located in the same contig and the contig sequences have sequences identical to those in the microbial species, the corresponding contigs needed to be removed from the reference genome. The gene sets annotated with *P. aeruginosa*, *P. salmonis*, Spodoptera moth adenovirus 1, and *Trichoplusia ni* TED virus descriptions were not located in the same contigs. Ten genes annotated with the *C. coronatus* descriptions were located in the contig; however, this contig did not hit any genome sequences of *C. coronatus*. Therefore, we did not remove any contigs from the *M. separata* reference genome sequences.

### 3.3. Immune-related genes in *M. separata*

To determine whether antimicrobial peptide (AMP) production systems function in *M. separata*, we searched for genes in the Toll and IMD pathways, which have been thoroughly investigated in *D. melanogaster* [19]. Using data sets of the Toll and IMD pathway sequences in FlyBase, we determined whether orthologs of the core components of the Toll and IMD pathways exist in *M. separata*. Orthologs of almost all core intracellular components of the Toll and IMD pathways were

found in *M. separata* (Figure 2). Although *IMD* and *MyD88* were not found by BLASTx (Supplementary Data 8), both genes were identified by BLASTp (Supplementary Data 9).

252  
253

**Extracellular components and sensor proteins**

PGRP:  
anno1.g18688.t1,anno1.g9300.t1,  
anno1.g9298.t1,anno1.g25239.t1,  
anno1.g25236.t1, anno1.g25238.t1,  
anno1.g25237.t1, anno1.g12592.t1,  
anno1.g14545.t2(low score)  
PGRP-LC:anno1.g18688.t1  
PGRP-LE:anno1.g18688.t1  
PGRP-SA:anno1.g9300.t1

GNBP:  
anno2.g19091.t1,anno1.g3877.t1,anno1.g3874.t1,anno1.g3878.t1,anno1.g3876.t1,anno1.g20324.t1,anno1.g20323.t1  
GNBP1:anno2.g19091.t1,GNBP3:anno1.g3878.t1  
ModSP:anno2.g17841.t1  
Grass:anno1.g9422.t1

**Intracellular components**

IMD\*: anno1.g708.t1  
IMD Pathway  
FADD:anno2.g12482.t1  
Toll Pathway  
Dredd:anno1.g24750.t1  
Diap2: anno2.g13458.t1  
IKK beta:anno1.g10261.t1  
Tak1:anno2.g23090.t1  
Tab2:anno2.g23220.t1  
Rel:anno1.g1369.t1

**Toll Pathway**

Myd88\*:anno1.g16865.t1  
Tube:anno1.g5504.t1  
Pelle:anno1.g19993.t1  
Cactus:anno1.g12785.t1  
Dif/Dorsal:anno2.g22161.t1,anno1.g1369.t1

**Antimicrobial peptide**

11 Defensin, 10 Cecropin, 6 Attacin, 5 Gloverin, 9 Moricin, and 5 Lebocin

**Figure 2.** Gene IDs of the orthologues consisting of Toll and IMD pathway based on BLASTx search. Number of genes of antimicrobial peptides were shown in bottom of this figure, and the gene IDs of the Orthologs of antimicrobial peptide are shown in Supplementary Data 10. Asterisks indicates that the orthologues were found by BLASTp search.

254  
255  
256  
257

Extracellular components in Toll pathways were searched [19]. Orthologues of *ModSP*, *Grass* and *Spaezle-processing enzyme (SPE)*, which are forming serine protease cascade between sensor proteins and Toll receptor complex, were found in *M. separata* genes while *Toll (Tl)* and *Spaezle (Spz)* (found by the BLASTp search) orthologs consisting of the Toll receptor complex were found (Figure 2). Sensor proteins of IMD and Toll pathway recognize the surface structures of the infected microbes and arouse signals to activate these pathways. PGRPs recognize the peptidoglycan structure derived from gram-positive or gram-negative bacteria. BLASTx results showed that nine genes in the *M. separata* genome were orthologs of PGRPs; orthologs of *PGRP-LE* and *-LC* (the IMD pathway) and *PGRP-SA* were found, and the genes showing minimum e-values among *M. separata* nine PGRPs in the BLASTx results were allocated as orthologue of each three PGRP (Figure 2). GNBP also acts as a sensor protein for fungi and gram-positive bacteria. Seven GNBP-encoding genes were found in the *M. separata* genome, whereas *D. melanogaster* harbored only three such genes. Orthologs of *GNBP1* and *GNBP3* showing minimum e-values among *M. separata* GNBP in the BLASTx results were allocated (Figure 2). Furthermore, the NJ phylogenetic trees of PGRPs and GNBP were constructed, and the domain analysis was performed by hmmerScan bundled in DoMosaics, and the ML phylogenetic trees of PGRP and GNBP were constructed (Supplemental Data 10). In the PGRP NJ tree, three *D. melanogaster* PGRPs and *M. separata* anno1.g18688.t1 (annotated PGRP-LC and LE by BLASTx results.) formed a clade while *M. separata* anno1.g9300.t1 belonged to another clade. The domain analysis of the PGRP showed that most of *M. separata* PGRPs possessed Amidase\_2 domain, which *D. melanogaster* PGRPs harbored. The structure of the ML tree of PGRP were different from that of the NJ tree of PGRP. In the ML tree, *D. melanogaster* PGRP-LC and *D. melanogaster* PGRP-SA formed a clade with anno1.g9298.t and anno1.g14545.t, respectively while *D. melanogaster* PGRP-LE did not form a clade with a single *M. separata* PGRP. In the GNBP NJ tree, anno1.g3877.t1 and anno1.g3876 formed the same clades of *D. melanogaster* GNBP1 and GNBP3, respectively, and both two *M. separata* GNBP possess CBM39 and Glyco\_hydro\_16 domain, which *D. melanogaster* GNBP harbored. The structures of the ML tree were different from that of the NJ tree. *D. melanogaster* GNBP1 formed a clade with two *M. separata* GNBP (anno1.g20324 and anno1.g20324), which

258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269  
270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281  
282  
283  
284

are considered as paralogous genes, while *D. melanogaster* GNBP3 did not form a clade with a single *M. separata* GNBP. Taken together, there are different relationships of orthologues between *D. melanogaster* and *M. separata* in the NJ and ML trees. Furthermore, there are inconsistencies between the BLASTx and phylogenetic tree results. The reasons of them might be because the methods of calculations of the sequence similarities were different, and the degrees of differences within PGRPs and GNBP sequences might not be so high, especially, in PGRPs (low bootstrap values in the ML tree). The functional analysis of these proteins of *M. separata* will be required in future research.

AMPs are effector molecules against invading microbes, and the induction of AMP genes is regulated mainly by the Toll and IMD pathways in *D. melanogaster* [49]. Defensin, cecropin, attacin, gloverin, moricin, and lebocin are the major AMP genes in insects [50]. Using *M. separata* gene dataset and annotation results (Supplementary Data 6), orthologs of the AMP genes in *M. separata* were searched (Supplementary Data 11). This is because most AMPs are very short, and orthologs cannot be found through BLAST alone. Although orthologs of defensin and cecropin were not found with low E-values either by BLASTp or hmmer domain search (data not shown), 11 defensins and 10 cecropins were found in the annotation data of *M. separata*. The 11 defensins were annotated with "Defensin" descriptions of several insect species such as *D. melanogaster* and *Aedes aegypti* ("UniGene-description" column in Supplemental Data 6), whereas the 10 cecropins were annotated with "Cecropin" descriptions of *M. sexta*. Six attacins, five gloverins (anno2.g16145.t1 was also annotated as attacin), nine moricins (seven moricins were also annotated as cecropins), and five lebocins were found in *M. separata*.

As described above, CTLs recognize a wide range of ligands [51,52]. Therefore, CTLs are involved in several immune reactions of insects by binding to the surfaces of not only the invaded microbes and organisms but also their own cells and tissues. CTLs possess more than one C-type lectin domain (CTLD) (also known as the "carbohydrate recognition domains"; CRD). Thus, CTL genes in *M. separata*, which were annotated as possessing CTLD (Pfam ID: PF00059), were searched. Among *M. separata* genes, 105 CTLs were identified (Supplemental Data 12). The results of the domain analysis by hmmerScan bundled in DoMosaics and sequences analysis of the 105 CTL are shown in Supplemental Data 13. Most CTL possessed two CTLDs classified as Dual CTLD, and 12 and 6 possessed single and more than three CTLDs, respectively (Table 5). Five CTLs possessed CTLD and other domains, which were classified as CTL-X. Furthermore, we determined whether the CTLs possessed signal peptide sequences and transmembrane domains. Of the 105 CTLs, 77 and 84 were predicted to possess signal peptides, using SignalP and DeepTMHMM, respectively (Table 5). This difference in the number of predicted signal peptides may be because of the difference in the prediction algorithms of the two software. These four CTLs possessed a transmembrane domain.

**Table 5.** Summary of CTL analyses

Total Number of CTLs	Single CTLD*	Dual CTLD (Immune-Lectin group or dual-CTLD type lectins)*	More than three CTLD	CTL-X group*	Signal peptide (SignalP)	Signal peptide (DeepTMHMM)	Transmembrane domain
105	12	81	7	5	77	84	4

\*Classifications are based on Xia et al., 2018 [52].

#### 4. Discussion

In this study, the reference genome sequences and gene dataset of the oriental armyworm, *M. separata*, were constructed. Using long-read and short-read genome data, genome sequences with high contig N50 values and good BUSCO scores were constructed through polishing and purging haplotig processes. After searching for repetitive regions and TEs in the genome sequences and masking such sequences, we obtained gene set data, using masked genome data and RNA-Seq data, some of which were retrieved from SRA. Consequently, the gene set was functionally annotated, and using amino acid sequence data of the gene set and the functional annotation data, orthologs of the IMD and Toll pathways as well as CTLs, which are involved in immune reactions in model insect species, were identified.

The contig N50 value of the reference genome we constructed was approximated at 2.7Mbp, and the results of the BUSCO analysis showed that over 98 % of BUSCO genes (core genes) were identified. These results suggest that the genome sequences data of *M. separata* possess sufficient quality as the reference genome in term of continuity and completeness of genome, which can be used for the various studies, especially genome or molecular research. The size of *M. separata* finished genome sequence was approximately 628 Mbp, as against these of *P. xylostella*, *T. ni*, *B. mori*, *S. litura*, *S. exigua*, and *M. sexta* that were 328 Mbp, 333Mbp, 460 Mbp, 438 Mbp, 419 Mbp, and 470 Mbp, respectively [2,9,12,13,15,17]. The genome size of butterflies (264 species) varied from 195 to 1262 Mbp [18]. Considering these results, the genome size of *M. separata* is larger than those of the other many lepidopteran species except those of butterflies, and but it is a reasonable in the range of some lepidopteran insects.

Further, it was found that approximately 48% of the *M. separata* genome contained repetitive regions or TEs, and retroelements were the major TEs, of which LINEs were the major components. In *B. mori*, approximately 46.5% of the genome contained repetitive regions or TEs, and more Class I retroelements than Class II DNA elements were detected, which is the same tendency as in the *M. separata* genome [9]. Additionally, SINEs consist of approximately 12% of the *B. mori* genome, and LINEs consist of approximately 17%, which are comparable numbers. However, approximately 32% of the *S. litura* genome and more Class I retroelements (about 11%) than Class II DNA elements (about 2%) were detected. LINEs and SINEs consist of about 8% and 2% of the *S. litura* genome, respectively [2]. In *A. mellifera*, a hymenopteran model species, the total repetitive region consists of approximately 11% of the genome, and Class II DNA elements are major components of the TEs, which are clearly different from lepidopteran families [53]. Taken together, these results suggest that the comprehensive tendency of *M. separata* was the same as the model lepidopteran species, and detailed status, especially the ratio of SINEs were different, which may be *M. separata*-specific features of TEs.

A total of 21,970 genes (24,452 CDSs) were predicted in the *M. separata* genome, whereas *B. mori*, *S. litura*, *S. exigua*, and *M. sexta* had 16,880 genes, 15,317 protein-coding genes, 18,477 transcripts, and 25,256 genes, respectively [2,9,12,17]. The differences among the species may be partially owing to the methods used to prepare the gene sets. Considering these results, the number of genes of *M. separata* may be reasonable for lepidopteran gene sets. The annotation ratios of *H. sapiens*, *M. musculus*, and *C. elegans* were lower than those of *T. castaneum*, *M. sexta*, and *B. mori*. The ratio of *D. melanogaster* to *A. mellifera*, a model insect species, was not very high. According to a phylogenetic tree constructed by Misof et al. [54], Hymenoptera branched at the earliest time, Coleoptera branched, and Lepidoptera and Diptera branched. Our ratios were not consistent with this tree structure. Although we could not determine the reasons underlying these differences, the gene sets could provide new evolutionary insights in such research fields.

Searching for the genes consisting of the Toll and IMD pathways revealed that intracellular components of these pathways are conserved, which has been observed in multiple insect species [19,55–57], suggesting that AMPs are regulated by the two signaling pathways in *M. separata*. The numbers of PGRPs and GNBP predicted to function as pattern recognition receptors (PRRs) for fungi and bacteria [58] were different from those in other insect species [19,55,57,59,60]. Nine PGRPs and seven GNBP were identified in the *M. separata* genome. Among these genes, the orthologs of PGRP-LC, -LE and -SA, GNBP1, and GNBP3 function as PRRs in *D. melanogaster*. However, functional analysis using some non-*Drosophila* species showed that the orthologs did not play a pivotal role in sensing microbes [60,61], which led to different signal transduction in the Toll and IMD pathways in *D. melanogaster*. The results suggest that *M. separata* has the Toll and IMD pathway systems, which have different signal transduction flows from *D. melanogaster*, as observed in *Plautia stali* and *T. castaneum* [60,62]. Over 40 AMP genes were identified in the *M. separata* genome and included typical AMP genes, such as cecropin, defensin, and attacin. However, some of the AMP genes were annotated without typical domain annotations in the Pfam database analysis. For example, the cecropin domain was not identified in the 10 cecropin genes annotated by the *M. sexta* gene set. Thus, further investigation is warranted to determine whether such AMP genes actually exist or are expressed, using RNA-Seq or quantitative reverse transcription-PCR analysis.

According to Xia et al., the number of CTL genes in insects ranges from 4 to 40 [52]. Among lepidoptera species, *B. mori* and *M. sexta* have 23 and 34 CTL genes, respectively, whereas *P. xylostella* has only seven genes. Compared to these species, *M. separata* has an outstandingly higher number of CTL genes. Notably, approximately 90% of CTLs have two CTLDs (or CRDs), which were categorized as dual CTLD type CTLs and more CTLD-type CTLs. Additionally, approximately >70% CTLs were found to have a signal peptide sequence, five CTLs were classified as CTL-X, and four CTLs had transmembrane domains, and these results are not different from those obtained for

other insects [52]. Based on these results, we assumed that the dual-CTLD-type lectin genes were duplicated in the *M. separata* genome. Many dual-CTLD-type lectins regulate the immune system of insects. Immunlectin II, a dual-CTLD-type lectin from *M. sexta*, binds to bacterial lipopolysaccharide with a CTLD [63]. In *B. mori*, dual-CTLD-type CTLs are required for nodule formation [64]. EPL, which enhances encapsulation, is also a dual-CTLD-type lectin [21]. Therefore, dual-CTLD-type lectins are expected to play more important roles in immunity in *M. separata* than in other insects. Recently, Sawa et al. reported that the parasitoid wasp *Cotesia cariyai* suppressed melanization and encapsulation of *M. separata* by manipulating the CTLs of both species, suggesting that CTLs are key factors for the success of parasitization [65]. To shed light on the well-controlled immune system and its interaction with various microorganisms and parasitoids, analyzing the diversified lectins and the molecules they recognize will be necessary.

## 5. Conclusions

We constructed the reference genome of *M. separata* (size: 682 Mbp) with sufficient qualities as the reference genome, in terms of continuity and completeness of core genes. In addition, 21,970 genes were predicted using genome sequence data, and functional annotations were performed. Using the gene set and annotation data, several immune-related genes were identified, providing new insights into the genomics and immunology of *M. separata*. We consider that the obtained reference genome data can promote studies on molecular biology of *M. separata* and comparative genomics of insects in general.

**Supplementary Materials:** All Supplemental materials are available in figshare. For the detail, See “Data Availability Statement” section.

**Author Contributions:** Conceptualization, K.Y., S.F. and H.B.; methodology, K.Y., A.J. and H.B.; validation, K.Y., A.J. and H.B.; formal analysis, K.Y. and H.B.; resources, S.F. and R.Z.; data curation, K.Y., S.F., A.J. and H.B.; writing—original draft preparation, K.Y.; writing—review and editing, K.Y., S.F., A.J. and H.B.; visualization, K.Y.; supervision, K.Y.; project administration, K.Y.; funding acquisition, K.Y. and S.F. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by JSPS KAKENHI Grant Number 16H06279 (PAGS) and funded by JSPS KAKENHI Grant Number 18K05669 to S.F. and 21K19126 to K.Y.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** All Supplemental data are available in figshare. DOI: 10.6084/m9.figshare.c.6192784.

**Supplemental Data 1** Sequence Read Archive accession IDs of raw sequence data in this study. DOI: 10.6084/m9.figshare.21082570.

**Supplemental Data 2** DDBJ/ENA/GenBank accession IDs and FASTA file IDs of assembled genome sequences. DOI: 10.6084/m9.figshare.21215255.

**Supplemental Data 3** Consensus sequences of transposable elements in *M. separata*. The consensus sequences of transposable elements in *M. separata* were constructed by RepeatModeler2. The sequence files are FASTA file and stk file from RepeatModeler2. DOI: 10.6084/m9.figshare.21230882

**Supplemental Data 4** Output files of RepeatMasker. Armyworm\_genome\_seq\_finished.fa.out shows the detailed status of all detected transposable elements (TEs) and their repetitive positions. Armyworm\_genome\_seq\_finished.fa.cat.gz shows the sequence comparison between all TEs and genome sequences. Armyworm\_genome\_seq\_finished.fa.masked.gz is the genome sequence in which repetitive and TE sequences are indicated in lowercase letters. See RepeatMasker for detailed explanation (<https://www.repeatmasker.org/>). DOI: 10.6084/m9.figshare.21231500.

**Supplemental Data 5** Data related to predicted genes, coding sequences and amino acid sequences in *M. separata* by Braker2. Gene.gtf contains structural data of predicted genes and transcripts. Gene\_cds.fasta and gene\_pep.fasta contain predicted coding sequences and amino acid sequences respectively. DOI: 10.6084/m9.figshare.21257601.

**Supplemental Data 6** Functional annotations of *M. separata* genes. Using the predicted amino acid sequences of *M. separata*, its genes were functionally annotated using Fanflow4Insects. The predicted amino acid sequences were compared with the gene set data of *Homo sapiens*, *Mus musculus*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Bombyx mori*, *Manduca sexta*, *Apis mellifera*, *Tribolium castaneum*, and UniProtKB/Swiss-Prot dataset. "pid" and "gene\_symbol" indicate hit protein IDs and gene symbols of each dataset, respectively. The protein domains were annotated using Pfam and HMMER. DOI: 10.6084/m9.figshare.21257625.

**Supplemental Data 7** BLASTp annotation results of the predicted gene set against the NCBI-nr database and count results of the number of species that annotate *M. separata* genes. Input \_pep.fasta.summary is the result of annotation of *M. separata* genes against the BLAST\_nr database. results.txt contains counting results for the number of species that annotate *M. separata* genes. DOI: 10.6084/m9.figshare.21386523.

**Supplemental Data 8** BLASTx results for genes in the Toll and IMD pathways. BLASTx was performed using either *D. melanogaster* transcript sequences consisting of the Toll (ID: FBgg0001194) or IMD (ID: FBgg0001059) pathway components in Flybase as query sequences and *M. separata*'s predicted amino acid sequences as subject sequences. DOI: 10.6084/m9.figshare.21325707.

**Supplemental Data 9** BLASTp results for genes in the Toll and IMD pathways. BLASTp was performed using either *D. melanogaster* amino acid sequences consisting of the Toll (ID: FBgg0001194) or IMD (ID: FBgg0001059) pathway components in Flybase as query sequences and *M. separata*'s predicted amino acid sequences as subject sequences. DOI: 10.6084/m9.figshare.21325740.

**Supplemental Data 10** Phylogenetic trees and domain analysis of PGRPs and GNBP of *M. separata* and *D. melanogaster* using DoMosaics and MEGA X. NJ phylogenetic trees of GNBP and PGRPs, plus the domain structure of each gene by DoMosaics, and ML phylogenetic trees by MEGA X with bootstrap values are shown in PGRP\_GNBP\_tree.pptx. PGRP.nwk and GNBP.nwk are the files used for the construction of a phylogenetic tree of DoMosaics. PGRP.domtree and GNBP.domtree are the raw files of the DoMosaics phylogenetic trees. Two mtsx files were session files of ML trees of PGRP and GNBP for MEGA X. Genbank accession IDs of *D. melanogaster* PGRP and GNBP are list in Accession\_IDs\_Dm\_GNBP\_PGRP.csv. DOI: 10.6084/m9.figshare.21709790.

**Supplemental Data 11** ID list of orthologs of the six antimicrobial peptide (AMP) genes. Orthologs of the six AMP genes (defensin, cecropin, attacin, gloverin, moricin, and lebocin) in *M. separata* were searched using the annotation results (Supplementary Data 6). The *M. separata* gene set IDs of the AMP orthologs are listed. DOI: 10.6084/m9.figshare.21346908.

**Supplemental Data 12** Domain and phylogenetic analysis of 105 C-type lectins identified in *M. separata*, using DoMosaics. CTL\_tree\_domain.jpg is the CTL tree developed using DoMosaics. C\_type\_lectin\_ID.txt shows the ID of C-type lectin genes in *M. separata*, and CTL.nwk is a new file for the construction of a phylogenetic tree. The CTL.domtree file is the raw file of the CTL\_tree\_domain.jpg. DOI: 10.6084/m9.figshare.21351501.

**Supplemental Data 13** SignalP results for signal peptide sequences and DeepTMHMM results for the transmembrane domain. output\_signalP.gff3 and SignalP\_results.txt are the output files of SignalP, and DTU\_DeepTMHMM\_1.0.15-results.zip is the output file of DeepTMHMM. For details, please refer to the references of the software. DOI: 10.6084/m9.figshare.21436560.

**Acknowledgments:** We deeply acknowledged Prof. Dr. Atsushi Toyoda at Advanced Genomics Center, The National Institute of Genetics, Mishima, Shizuoka, Japan for the sequencing the prepared gDNA samples.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Goergen, G.; Kumar, P.L.; Sankung, S.B.; Togola, A.; Tamò, M. First Report of Outbreaks of the Fall Armyworm Spodoptera Frugiperda (J E Smith) (Lepidoptera, Noctuidae), a New Alien Invasive Pest in West and Central Africa. *PloS One* **2016**, *11*, e0165632, doi:10.1371/journal.pone.0165632.

2. Cheng, T.; Wu, J.; Wu, Y.; Chilukuri, R.V.; Huang, L.; Yamamoto, K.; Feng, L.; Li, W.; Chen, Z.; Guo, H.; et al. Genomic Adaptation to Polyphagy and Insecticides in a Major East Asian Noctuid Pest. *Nat. Ecol. Evol.* **2017**, *1*, 1747–1756, doi:10.1038/s41559-017-0314-4. 496

3. Sharma, H.C.; Davies, J.C. The Oriental Armyworm, *Mythimna Separata* (Wlk.). Distribution, Biology and Control: A Literature Review. *Orient. Armyworm Myth. Sep. Wlk Distrib. Biol. Control Lit. Rev.* **1983**. 497

4. Hiruma, K.; Riddiford, L.M. Developmental Expression of MRNAs for Epidermal and Fat Body Proteins and Hormonally Regulated Transcription Factors in the Tobacco Hornworm, *Manduca Sexta*. *J. Insect Physiol.* **2010**, *56*, 1390–1395, doi:10.1016/j.jinsphys.2010.03.029. 498

5. Kanost, M.R.; Jiang, H.; Yu, X.-Q. Innate Immune Responses of a Lepidopteran Insect, *Manduca Sexta*. *Immunol. Rev.* **2004**, *198*, 97–105, doi:10.1111/j.0105-2896.2004.0121.x. 499

6. Mita, K.; Kasahara, M.; Sasaki, S.; Nagayasu, Y.; Yamada, T.; Kanamori, H.; Namiki, N.; Kitagawa, M.; Yamashita, H.; Yasukochi, Y. The Genome Sequence of Silkworm, *Bombyx Mori*. *DNA Res.* **2004**, *11*, 27–35. 500

7. Xia, Q.; Zhou, Z.; Lu, C.; Cheng, D.; Dai, F.; Li, B.; Zhao, P.; Zha, X.; Cheng, T.; Chai, C. A Draft Sequence for the Genome of the Domesticated Silkworm (*Bombyx Mori*). *Science* **2004**, *306*, 1937–1940. 501

8. The International Silkworm Genome Consortium The Genome of a Lepidopteran Model Insect, the Silkworm *Bombyx Mori*. *Insect Biochem. Mol. Biol.* **2008**, *38*, 1036–1045, doi:10.1016/j.ibmb.2008.11.004. 502

9. Kawamoto, M.; Jouraku, A.; Toyoda, A.; Yokoi, K.; Minakuchi, Y.; Katsuma, S.; Fujiyama, A.; Kiu-chi, T.; Yamamoto, K.; Shimada, T. High-Quality Genome Assembly of the Silkworm, *Bombyx Mori*. *Insect Biochem. Mol. Biol.* **2019**, *107*, 53–62. 503

10. Gouin, A.; Bretaudeau, A.; Nam, K.; Gimenez, S.; Aury, J.-M.; Duvic, B.; Hilliou, F.; Durand, N.; Montagné, N.; Darboux, I.; et al. Two Genomes of Highly Polyphagous Lepidopteran Pests (*Spodoptera Frugiperda*, *Noctuidae*) with Different Host-Plant Ranges. *Sci. Rep.* **2017**, *7*, 11816, doi:10.1038/s41598-017-10461-4. 504

11. Kakumani, P.K.; Malhotra, P.; Mukherjee, S.K.; Bhatnagar, R.K. A Draft Genome Assembly of the Army Worm, *Spodoptera Frugiperda*. *Genomics* **2014**, *104*, 134–143, doi:10.1016/j.ygeno.2014.06.005. 505

12. Simon, S.; Breeschoten, T.; Jansen, H.J.; Dirks, R.P.; Schranz, M.E.; Ros, V.I.D. Genome and Transcriptome Analysis of the Beet Armyworm *Spodoptera Exigua* Reveals Targets for Pest Control. *G3 Bethesda Md* **2021**, *11*, jkab311, doi:10.1093/g3journal/jkab311. 506

13. Ward, C.M.; Perry, K.D.; Baker, G.; Powis, K.; Heckel, D.G.; Baxter, S.W. A Haploid Diamondback Moth (*Plutella Xyllostella* L.) Genome Assembly Resolves 31 Chromosomes and Identifies a Diamide Resistance Mutation. *Insect Biochem. Mol. Biol.* **2021**, *138*, 103622, doi:10.1016/j.ibmb.2021.103622. 507

14. You, M.; Yue, Z.; He, W.; Yang, X.; Yang, G.; Xie, M.; Zhan, D.; Baxter, S.W.; Vasseur, L.; Gurr, G.M.; et al. A Heterozygous Moth Genome Provides Insights into Herbivory and Detoxification. *Nat. Genet.* **2013**, *45*, 220–225, doi:10.1038/ng.2524. 508

15. Chen, W.; Yang, X.; Tetreau, G.; Song, X.; Coutu, C.; Hegedus, D.; Blissard, G.; Fei, Z.; Wang, P. A High-Quality Chromosome-Level Genome Assembly of a Generalist Herbivore, *Trichoplusia Ni*. *Mol. Ecol. Resour.* **2019**, *19*, 485–496, doi:10.1111/1755-0998.12966. 509

16. Kanost, M.R.; Arrese, E.L.; Cao, X.; Chen, Y.-R.; Chellapilla, S.; Goldsmith, M.R.; Grosse-Wilde, E.; Heckel, D.G.; Herndon, N.; Jiang, H.; et al. Multifaceted Biological Insights from a Draft Genome Sequence of the Tobacco Hornworm Moth, *Manduca Sexta*. *Insect Biochem. Mol. Biol.* **2016**, *76*, 118–147, doi:10.1016/j.ibmb.2016.07.005. 510

17. Gershman, A.; Romer, T.G.; Fan, Y.; Razaghi, R.; Smith, W.A.; Timp, W. De Novo Genome Assembly of the Tobacco Hornworm Moth (*Manduca Sexta*). *G3 GenesGenomesGenetics* **2021**, *11*, jkaa047, 538  
doi:10.1093/g3journal/jkaa047. 539

18. Liu, G.; Chang, Z.; Chen, L.; He, J.; Dong, Z.; Yang, J.; Lu, S.; Zhao, R.; Wan, W.; Ma, G.; et al. Genome Size Variation in Butterflies (Insecta, Lepidoptera, Papilionoidea): A Thorough Phylogenetic Comparison. *Syst. Entomol.* **2020**, *45*, 571–582, doi:10.1111/syen.12417. 540

19. Lemaitre, B.; Hoffmann, J. The Host Defense of *Drosophila Melanogaster*. *Annu. Rev. Immunol.* **2007**, 541  
*25*, 697–743, doi:10.1146/annurev.immunol.25.022106.141615. 542

20. Oda, Y.; Matsumoto, H.; Kurakake, M.; Ochiai, M.; Ohnishi, A.; Hayakawa, Y. Adaptor Protein Is Essential for Insect Cytokine Signaling in Hemocytes. *Proc. Natl. Acad. Sci.* **2010**, *107*, 15862–15867, 543  
doi:10.1073/pnas.1003785107. 544

21. Ishihara, T.; Maruyama, Y.; Furukawa, S. Gene Expression and Molecular Characterization of a Novel C-Type Lectin, Encapsulation Promoting Lectin (EPL), in the Rice Armyworm, *Mythimna Separata*. *Insect Biochem. Mol. Biol.* **2017**, *89*, 51–57, doi:10.1016/j.ibmb.2017.08.008. 545

22. Song, Z.; Tian, M.; Dong, Y.; Ren, C.; Du, Y.; Hu, J. The C-Type Lectin IML-10 Promotes Hemocytic Encapsulation by Enhancing Aggregation of Hemocytes in the Asian Corn Borer *Ostrinia Furnacalis*. *Insect Biochem. Mol. Biol.* **2020**, *118*, 103314, doi:10.1016/j.ibmb.2020.103314. 546

23. Li, H.; Liu, F.-F.; Fu, L.-Q.; Liu, Z.; Zhang, W.-T.; Wang, Q.; Rao, X.-J. Identification of 35 C-Type Lectins in the Oriental Armyworm, *Mythimna Separata* (Walker). *Insects* **2021**, *12*, 559, doi:10.3390/insects12060559. 547

24. Harvey, J.A.; Poelman, E.H.; Tanaka, T. Intrinsic Inter- and Intraspecific Competition in Parasitoid Wasps. *Annu. Rev. Entomol.* **2013**, *58*, 333–351, doi:10.1146/annurev-ento-120811-153622. 548

25. Suzuki, M.; Tanaka, T. Development of *Meteorus Pulchricornis* and Regulation of Its Noctuid Host, *Pseudaletia Separata*. *J. Insect Physiol.* **2007**, *53*, 1072–1078, doi:10.1016/j.jinsphys.2007.06.006. 549

26. Suzuki, M.; Miura, K.; Tanaka, T. The Virus-like Particles of a Braconid Endoparasitoid Wasp, *Meteorus Pulchricornis*, Inhibit Hemocyte Spreading in Its Noctuid Host, *Pseudaletia Separata*. *J. Insect Physiol.* **2008**, *54*, 1015–1022, doi:10.1016/j.jinsphys.2008.03.013. 550

27. Yamashita, K.; Zhang, K.; Ichiki, R.T.; Nakamura, S.; Furukawa, S. Novel Host Immune Evasion Strategy of the Endoparasitoid *Drino Inconspicuoides*. *Bull. Entomol. Res.* **2019**, *109*, 643–648, 551  
doi:10.1017/S0007485318001049. 552

28. Schwier, N.; Zhang, K.; Nakamura, S.; Furukawa, S. Larvae of the Tachinid Fly, *Drino Inconspicuoides* (Diptera: Tachinidae), Suppress Melanization in Host Lepidopteran Insects. *J. Asia-Pac. Entomol.* **2021**, *24*, 1050–1054, doi:10.1016/j.aspen.2021.09.012. 553

29. Yokoi, K.; Sano, T.; Suzuki, M.; Tanaka, T.; Minakuchi, C.; Miura, K. The Major Constituents of the Venom Gland of a Braconid Endoparasitoid, *Meteorus Pulchricornis* (Hymenoptera: Braconidae). *Appl. Entomol. Zool.* **2017**, *52*, 271–285. 554

30. Yokoi, K.; Kato, Y.; Suzuki, M.; Miura, K. Molecular Cloning and Functional Analyses of an Adhesion Molecule, Neuroglian, in *Mythimna Separata* (Lepidoptera: Noctuidae). *Eur. J. Entomol.* **2018**, *115*. 555

31. Kamezaki, M.; Yokoi, K.; Miura, K. RNA Interference Mediated Knockdown of an Inhibitor of Apoptosis Protein Induces Apoptosis in *Mythimna Separata* (Lepidoptera: Noctuidae). *Eur. J. Entomol.* **2018**, 556  
*115*. 557

32. Takabayashi, J.; Noda, T.; Takahashi, S. Effect of Kairomones in the Host Searching Behavior of Apanteles Kariyai WATANABE (Hymenoptera : Braconidae), a Parasitoid of the Common Armyworm, Pseudaletia Separata WALKER(Lepidoptera : Noctuidae). : I. Presence of Arresting Stimulants Produced by the Host Larvae. *Appl. Entomol. Zool.* **1985**, *20*, 484–489, doi:10.1303/aez.20.484. 579

33. Kolmogorov, M.; Yuan, J.; Lin, Y.; Pevzner, P.A. Assembly of Long, Error-Prone Reads Using Repeat Graphs. *Nat. Biotechnol.* **2019**, *37*, 540–546, doi:10.1038/s41587-019-0072-8. 580

34. Li, H. Minimap2: Pairwise Alignment for Nucleotide Sequences. *Bioinformatics* **2018**, *34*, 3094–3100, doi:10.1093/bioinformatics/bty191. 581

35. Danecek, P.; Bonfield, J.K.; Liddle, J.; Marshall, J.; Ohan, V.; Pollard, M.O.; Whitwham, A.; Keane, T.; McCarthy, S.A.; Davies, R.M.; et al. Twelve Years of SAMtools and BCFtools. *GigaScience* **2021**, *10*, giab008, doi:10.1093/gigascience/giab008. 582

36. Walker, B.J.; Abeel, T.; Shea, T.; Priest, M.; Abouelliel, A.; Sakthikumar, S.; Cuomo, C.A.; Zeng, Q.; Wortman, J.; Young, S.K.; et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE* **2014**, *9*, e112963, doi:10.1371/journal.pone.0112963. 583

37. Manni, M.; Berkeley, M.R.; Seppey, M.; Simão, F.A.; Zdobnov, E.M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.* **2021**, *38*, 4647–4654, doi:10.1093/molbev/msab199. 584

38. Flynn, J.M.; Hubley, R.; Goubert, C.; Rosen, J.; Clark, A.G.; Feschotte, C.; Smit, A.F. RepeatModeler2 for Automated Genomic Discovery of Transposable Element Families. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, 9451–9457, doi:10.1073/pnas.1921046117. 585

39. Smit, A.; Hubley, R.; Green, P. RepeatMasker Open-4.0. **2013**. 586

40. Kim, D.; Paggi, J.M.; Park, C.; Bennett, C.; Salzberg, S.L. Graph-Based Genome Alignment and Genotyping with HISAT2 and HISAT-Genotype. *Nat. Biotechnol.* **2019**, *37*, 907–915, doi:10.1038/s41587-019-0201-4. 587

41. Brúna, T.; Hoff, K.J.; Lomsadze, A.; Stanke, M.; Borodovsky, M. BRAKER2: Automatic Eukaryotic Genome Annotation with GeneMark-EP+ and AUGUSTUS Supported by a Protein Database. *NAR Genomics Bioinformatics* **2021**, *3*, lqaa108, doi:10.1093/nargab/lqaa108. 588

42. Gabriel, L.; Hoff, K.J.; Brúna, T.; Borodovsky, M.; Stanke, M. TSEBRA: Transcript Selector for BRAKER. *BMC Bioinformatics* **2021**, *22*, 566, doi:10.1186/s12859-021-04482-0. 589

43. Bono, H.; Sakamoto, T.; Kasukawa, T.; Tabunoki, H. Systematic Functional Annotation Workflow for Insects. *Insects* **2022**, *13*, 586, doi:10.3390/insects13070586. 590

44. Moore, A.D.; Held, A.; Terrapon, N.; Weiner, J., 3rd; Bornberg-Bauer, E. DoMosaics: Software for Domain Arrangement Visualization and Domain-Centric Analysis of Proteins. *Bioinformatics* **2014**, *30*, 282–283, doi:10.1093/bioinformatics/btt640. 591

45. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **2018**, *35*, 1547–1549. 592

46. Teufel, F.; Almagro Armenteros, J.J.; Johansen, A.R.; Gíslason, M.H.; Pihl, S.I.; Tsirigos, K.D.; Winther, O.; Brunak, S.; von Heijne, G.; Nielsen, H. SignalP 6.0 Predicts All Five Types of Signal Peptides Using Protein Language Models. *Nat. Biotechnol.* **2022**, *40*, 1023–1025, doi:10.1038/s41587-021-01156-3. 593

47. Hallgren, J.; Tsirigos, K.D.; Pedersen, M.D.; Armenteros, J.J.A.; Marcatili, P.; Nielsen, H.; Krogh, A.; Winther, O. DeepTMHMM Predicts Alpha and Beta Transmembrane Proteins Using Deep Neural Networks 2022, 2022.04.08.487609. 594

48. Roach, M.J.; Schmidt, S.A.; Borneman, A.R. Purge Haplotigs: Allelic Contig Reassignment for Third- 621  
Gen Diploid Genome Assemblies. *BMC Bioinformatics* **2018**, *19*, 460, doi:10.1186/s12859-018-2485-7. 622

49. De Gregorio, E.; Spellman, P.T.; Rubin, G.M.; Lemaitre, B. Genome-Wide Analysis of the Drosophila 623  
Immune Response by Using Oligonucleotide Microarrays. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 12590– 624  
12595, doi:10.1073/pnas.221458698. 625

50. Yi, H.-Y.; Chowdhury, M.; Huang, Y.-D.; Yu, X.-Q. Insect Antimicrobial Peptides and Their Applications. 626  
*Appl. Microbiol. Biotechnol.* **2014**, *98*, 5807–5822, doi:10.1007/s00253-014-5792-6. 627

51. Brown, G.D.; Willment, J.A.; Whitehead, L. C-Type Lectins in Immunity and Homeostasis. *Nat. Rev. 628  
Immunol.* **2018**, *18*, 374–389, doi:10.1038/s41577-018-0004-8. 629

52. Xia, X.; You, M.; Rao, X.-J.; Yu, X.-Q. Insect C-Type Lectins in Innate Immunity. *Dev. Comp. Immunol.* 630  
**2018**, *83*, 70–79, doi:10.1016/j.dci.2017.11.020. 631

53. Yokoi, K.; Kimura, K.; Bono, H. Revealing Landscapes of Transposable Elements in *Apis* Species by 632  
Meta-Analysis. *Insects* **2022**, *13*, 698, doi:10.3390/insects13080698. 633

54. Misof, B.; Liu, S.; Meusemann, K.; Peters, R.S.; Donath, A.; Mayer, C.; Frandsen, P.B.; Ware, J.; 634  
Flouri, T.; Beutel, R.G.; et al. Phylogenomics Resolves the Timing and Pattern of Insect Evolution. *Science* 635  
**2014**, *346*, 763–767, doi:10.1126/science.1257570. 636

55. Zou, Z.; Evans, J.D.; Lu, Z.; Zhao, P.; Williams, M.; Sumathipala, N.; Hetru, C.; Hultmark, D.; Jiang, 637  
H. Comparative Genomic Analysis of the *Tribolium* Immune System. *Genome Biol.* **2007**, *8*, R177, 638  
doi:10.1186/gb-2007-8-8-r177. 639

56. Yokoi, K.; Uchiyama, H.; Wakamiya, T.; Yoshiyama, M.; Takahashi, J.-I.; Nomura, T.; Furukawa, T.; 640  
Yajima, S.; Kimura, K. The Draft Genome Sequence of the Japanese Honey Bee, *Apis Cerana Japonica* 641  
(Hymenoptera: Apidae). *Eur. J. Entomol.* **2018**, *115*. 642

57. Evans, J.D.; Aronstein, K.; Chen, Y.P.; Hetru, C.; Imler, J.-L.; Jiang, H.; Kanost, M.; Thompson, G.J.; 643  
Zou, Z.; Hultmark, D. Immune Pathways and Defence Mechanisms in Honey Bees *Apis Mellifera*. *Insect 644  
Mol. Biol.* **2006**, *15*, 645–656, doi:10.1111/j.1365-2583.2006.00682.x. 645

58. Stokes, B.A.; Yadav, S.; Shokal, U.; Smith, L.C.; Eleftherianos, I. Bacterial and Fungal Pattern Recognition 646  
Receptors in Homologous Innate Signaling Pathways of Insects and Mammals. *Front. Microbiol.* 647  
**2015**, *6*. 648

59. Bao, Y.-Y.; Qu, L.-Y.; Zhao, D.; Chen, L.-B.; Jin, H.-Y.; Xu, L.-M.; Cheng, J.-A.; Zhang, C.-X. The 649  
Genome- and Transcriptome-Wide Analysis of Innate Immunity in the Brown Planthopper, *Nilaparvata* 650  
Lugens. *BMC Genomics* **2013**, *14*, 160, doi:10.1186/1471-2164-14-160. 651

60. Nishide, Y.; Kageyama, D.; Yokoi, K.; Jouraku, A.; Tanaka, H.; Futahashi, R.; Fukatsu, T. Functional 652  
Crosstalk across IMD and Toll Pathways: Insight into the Evolution of Incomplete Immune Cascades. *Proc. 653  
R. Soc. B Biol. Sci.* **2019**, *286*, 20182207, doi:10.1098/rspb.2018.2207. 654

61. Koyama, H.; Kato, D.; Minakuchi, C.; Tanaka, T.; Yokoi, K.; Miura, K. Peptidoglycan Recognition 655  
Protein Genes and Their Roles in the Innate Immune Pathways of the Red Flour Beetle, *Tribolium Castaneum*. *J. Invertebr. Pathol.* **2015**, *132*, 86–100, doi:10.1016/j.jip.2015.09.003. 656

62. Yokoi, K.; Koyama, H.; Minakuchi, C.; Tanaka, T.; Miura, K. Antimicrobial Peptide Gene Induction, 657  
Involvement of Toll and IMD Pathways and Defense against Bacteria in the Red Flour Beetle, *Tribolium* 658  
*Castaneum*. *Results Immunol.* **2012**, *2*, 72–82, doi:10.1016/j.rim.2012.03.002. 659

63. Yu, X.-Q.; Kanost, M.R. *Manduca Sexta* Lipopolysaccharide-Specific Immulectin-2 Protects Larvae 660  
from Bacterial Infection. *Dev. Comp. Immunol.* **2003**, *27*, 189–196, doi:10.1016/s0145-305x(02)00099-x. 661

662

---

64. Suzuki, A.; Yoshizawa, Y.; Tanaka, S.; Kitami, M.; Sato, R. Extra- and Intracellular Signaling Pathways Regulating Nodule Formation by Hemocytes of the Silkworm, *Bombyx Mori*(Lepidoptera: Bombycidae). *J. Insect Biotechnol. Sericology* **2011**, *80*, 2\_049-2\_056, doi:10.11416/jibs.80.2\_049. 663

65. Sawa, T.; Tanaka, T.; Kato, Y.; Nakamatsu, Y. Cky811 Protein Expressed by Polydnavirus and Venom Gland of Cotesia Kariyai Regulates the Host *Mythimna Separata* Larvae Immune Response Function of C-Type Lectin Responsible for Foreign Substance Recognition Which Suppresses Its Melanization and Encapsulation. *Arch. Insect Biochem. Physiol.* **2021**, *107*, e21786, doi:10.1002/arch.21786. 664

665

666

667

668

669

670