Article

Meta-analysis of the public RNA-Seq data of the Western honeybee *Apis mellifera* to construct reference transcriptome data

Kakeru Yokoi[1], Takeshi Wakamiya[2], Hidemasa Bono[2, 3,*]

[1] Insect Design Technology Module, Division of Insect Advanced Technology, Institute of Agrobiological Sciences, National Agriculture and Food Research Organization (NARO), 1-2 Owashi, Tsukuba 305-8634, Ibaraki, Japan; yokoi123@affrc.go.jp
[2] Laboratory of Genome Informatics, Graduate School of Integrated Sciences for Life, Hiroshima University, 3-10-23 Kagamiyama, Higashi-Hiroshima City, Hiroshima 739-0046, Japan; takeshi.waka38@gmail.com
[3] Laboratory of BioDX, Genome Editing Innovation Center, Hiroshima University, 3-10-23 Kagamiyama, Higashi-Hiroshima City, Hiroshima 739-0046, Japan; bonohu@hiroshima-u.ac.jp
* Correspondence: bonohu@hiroshima-u.ac.jp; Tel. +81-82-424-4013

Simple Summary: The honeybee is an economically and ecologically essential insect species and a critical social behavior model. It is vital to honey production and pollinates. It has also been studied extensively as a model social insect species in entomological research. The Western honeybee (*Apis mellifera*) is a representative species of honeybees. The reference transcriptome data are required to perform the meta-analyses using public RNA-Seq data and transcriptome analysis using newly sequenced RNA-Seq data. The reference transcriptome sequences was constructed, which consisted of 149,685 transcripts, and 194,174 predicted amino acid sequences. Half the predicted protein sequences were functionally annotated using protein sequence datasets for several model and insect species. We performed meta-analysis to search for novel immune response-related transcripts. We identified 3–20 transcripts that were upregulated or downregulated in response to both viral and bacterial infections. However, the immune-related transcripts were not included among them. Autophagy related protein 3 was downregulated by viral infections. The results suggested that autophagy was implicated in the viral immune responses, and autophagy was not regulated by the canonical immune pathways in *A. mellifera*. The results of the meta-analysis suggested that the reference transcriptome data could be used for transcriptome analysis in *A. mellifera.*

Abstract: The Western honeybee (*Apis mellifera*) is valuable in biological research and agriculture. Its genome sequence was published before those for other insect species. RNA-Seq data for *A. mellifera* have been applied in several recently published studies. Nevertheless, these data have not been prepared for use in subsequent meta-analyses. To promote *A. mellifera* transcriptome analysis, we constructed reference transcriptome data using the reference genome sequence and RNA-Seq data curated from about 1,000 runs of public databases. The new reference transcriptome data construct comprised 149,685 transcripts, and 194,174 protein sequences were predicted. About 50–60% of the predicted protein sequences were functionally annotated using the protein sequence data for several model and insect species. Novel candidate immune-related transcripts were searched by meta-analysis using immune response-related RNA-Seq and reference transcriptome data. Three to 20 candidate transcripts including *autophagy related protein 3* were upregulated or downregulated in response to both viral and bacterial infections. The constructed reference transcriptome data may facilitate future transcriptome analyses of *A. mellifera.*

Keywords: *Apis mellifera*; honeybee; meta-analysis; public RNA sequencing data; reference transcriptome

1. Introduction

The honeybee (Hymenoptera; Apidae; *Apis* spp.) is an economically and ecologically important insect species and a key social behavior model [1]. It is vital to honey production and pollinates both domesticated crops and wild plants. It has also been studied extensively as a model social insect species in entomological research. *Apis mellifera* (Western honeybee) is a representative *Apis* species that is widely applied in biological research, as a pollinator, and for honey production. In view of its utility and versatility, its reference genome sequence data were published in 2006 without the use of next-generation sequencing (NGS) [2]. The reference genome sequence data were continuously updated and chromosome-level genome assembly data were finally published in 2019 following advances in DNA sequencing technology [3]. Certain investigations using genomic and transcriptomic data for *A. mellifera* disclosed novel insights based on molecular and genetic evidence ([4]). Hence, the amount of *A. mellifera* transcriptome data generated by RNA sequencing (RNA-Seq) has increased in public databases.

The foregoing findings suggest that RNA-Seq and transcriptome data as well as studies using the RNA-Seq data of *A. mellifera* are accumulating. We previously found > 80 transcriptomic expression datasets for *A. mellifera* in the integrated transcriptome database known as All of Gene Expression (AOE) [5]. We also discovered several studies using *A. mellifera* RNA-Seq published over the past two years [6–9]. Future studies should use detailed transcriptomic data including isoform sequence data for single genes, non-coding RNA sequences, and annotations of each isoform. We constructed chromosome-level genome assembly data and gene sets [10] and reference transcriptome data [11] for the domestic silkworm (*Bombyx mori*) and demonstrated that studies could be conducted utilizing reference transcriptome data [12,13]. Considering the foregoing results, reference transcriptomic data for *Apis mellifera* should be constructed to foster *A. mellifera* research using transcriptomic analysis.

Meta-analysis is a promising method to use and combine the data and observations of different studies to obtain novel biological insights. We previously performed a meta-analysis using RNA-Seq or genomic data derived from various studies and available in public databases. We used public RNA-Seq data related to hypoxia to identify new candidate genes induced by hypoxic stimulation [14]. New candidate genes related to oxidative stress and hypoxia were identified using public RNA-Seq data associated with these pathological processes. Several of the genes identified were annotated as cell cycle functions [15]. Landscapes of the transposable elements (TEs) in *Apis* species were constructed using public genome sequence data for multiple *Apis* species. Most of the *Apis* TEs identified belonged to only a few families [16]. In addition, Doublet et al., showed that genes or molecular pathways in *A. mellifera* which were responded to multiple pathogens were identified using transcriptome dataset of RNA-Seq, microarray, or tilling array from diverse tissues [17]. The results of earlier studies suggested that meta-analyses involving sequence data in public database are highly effective research methods.

In the present study, we used a meta-analysis to construct reference transcriptome sequences comprising *A. mellifera* RNA-Seq data and chromosome-level genome assembly [3] from public databases. We then functionally annotated the constructed transcriptome sequences. We quantified the transcriptome expression levels using the public RNA-Seq data to show the usefulness of the reference data. We extracted the sample RNA from *A. mellifera* subjected to several microbial pathogens and identified candidate genes involved in the host immune response (Figure 1).

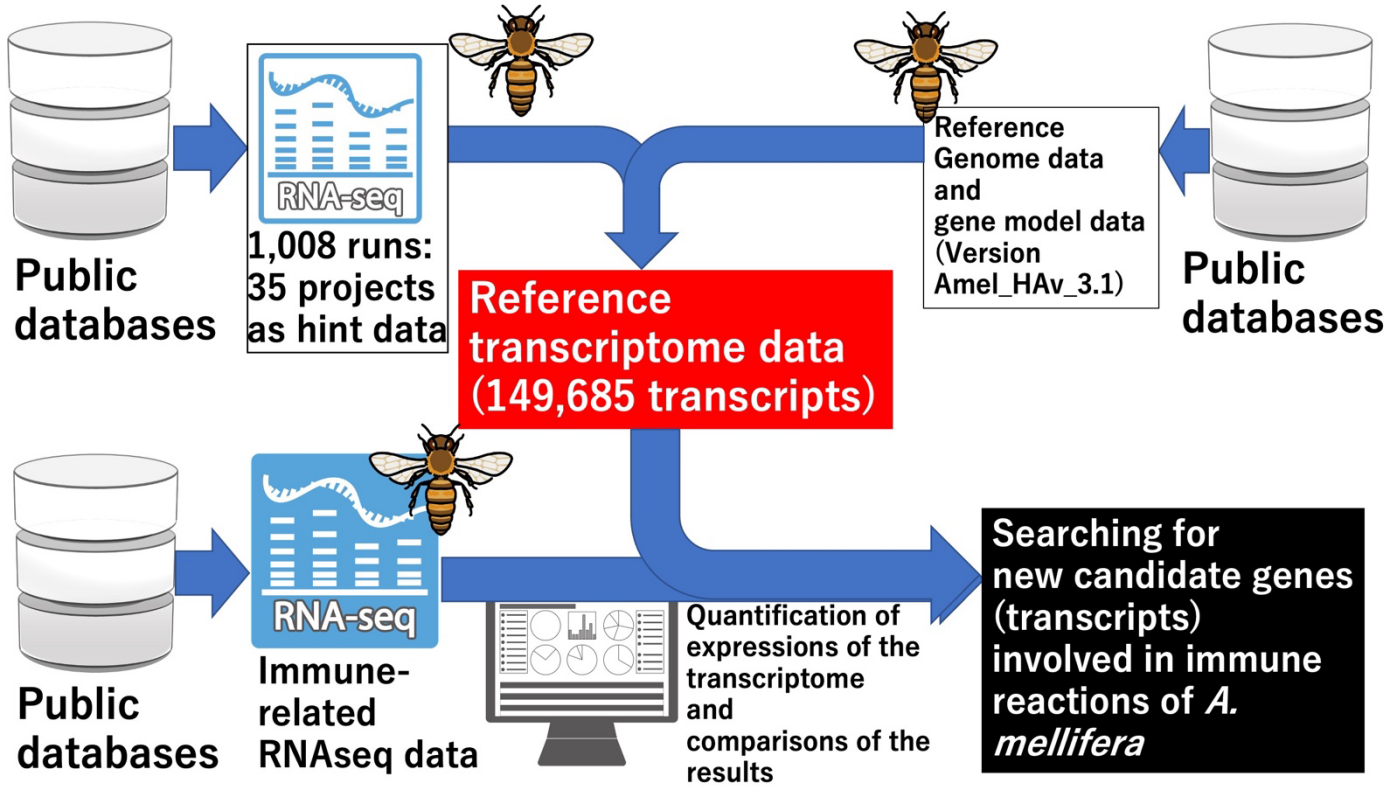

Figure 1. Workflow of the data analyses performed in the present study. To construct the reference transcriptome data, 1,008 RNA-Seq data from 35 projects were mapped to the reference genome data used as hint data. The reference transcriptome data were constructed using a combination of reference genome, genome model, and hint data and they were functionally annotated with Fanflow4Insects [18]. We retrieved and curated immune response-related RNA-Seq data in the public database to search novel candidate transcripts implicated in these processes. We quantified the expression levels of the RNA-Seq data transcripts using the reference transcriptome data. We searched new candidate transcripts by comparing the expression data. Images were obtained from TogoTV (© 2016 DBCLS TogoTV).

2. Materials and Methods

2.1. RNA-Seq data retrieval from the public database

The NCBI Gene Expression Omnibus database (GEO, https://www.ncbi.nlm.nih.gov/geo/; accessed August 1, 2022) was used to compile a list of *A. mellifera* RNA-Seq data. Each selected sequence read archive (SRA) file was downloaded using the 'prefetch' command in sra-tools v. 2.9.6 (https://github.com/ncbi/sra-tools; accessed August 1, 2022) and converted to FASTQ file using the 'fasterq-dump' command. The FASTQ files were compressed and trimmed using the default options in Trim Galore! v. 0.6.6 (https://github.com/FelixKrueger/TrimGalore; accessed August 1, 2022).

2.2. Constructing and evaluating the reference transcriptome sequence data

The trimmed RNA-Seq reads were mapped to the *A. mellifera* genome (v. Amel_HAv_3.1; GenBank assembly accession ID No. GCA_003254395.2) [3] using HISAT2 v. 2.2.1 (https://daehwankimlab.github.io/hisat2/; accessed August 1, 2022) [19]. The output SAM files were converted to BAM files with Samtools v. 1.12 (https://sourceforge.net/projects/samtools/files/samtools/1.12/; accessed August 1, 2022) [20] and the mapped BAM file data were used to assemble the reference transcriptome with the reference *A. mellifera* gene model GCF_003254395.2_Amel_HAv3.1_genomic.gtf.gz (https://ftp.ncbi.nlm.nih.gov/genomes/all/annotation_releases/7460/104/GCF_003254395.2_Amel_HAv3.1/GCF_003254395.2_Amel_HAv3.1_genomic.gtf.gz in NCBI Refseq accessed August 1, 2022) in a GTF file via StringTie v. 2.1.7

(https://github.com/gpertea/stringtie/releases; accessed August 1, 2022) [19]. The reference transcriptome sequence data in the GTF file were constructed by merging the transcript data of each RNA-Seq via the 'stringtie -merge' command. The GTF file of the reference transcriptome was converted into a FASTA file with gffread v. 0.12.1 (https://github.com/gpertea/gffread/releases; accessed August 1, 2022) [21]. The amino acid sequences of the reference transcriptome were predicted with TransDecorder v. 5.5.0 (https://github.com/TransDecoder/TransDecoder/wiki; accessed August 1, 2022) [22]. GffCompare v. 0.11.2 (https://anaconda.org/bioconda/gffcompare; accessed August 1, 2022) was used to compare the transcriptome data against the reference gene model [21]. The amino acids of the transcriptome data were compared against those for the reference gene model (GCF_003254395.2_Amel_HAv3.1_protein.faa; available in NCBI refseq) in BLAST v. 2.9.0+ (https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/; accessed August 1, 2022) using blastp with e-value < 1e-10 [23].

2.3. Reference transcriptome sequence annotations

The constructed transcriptome sequences were annotated via the functional annotation workflow Fanflow4Insects using the amino acid sequences predicted by TransDecoder (https://github.com/TransDecoder/TransDecoder; accessed August 1, 2022) [18]. The protein-level functional annotation consisted of assigning the top hit genes from comprehensively annotated organisms using the sequence similarity search program in global alignment (GGSEARCH) (https://fasta.bioch.virginia.edu/; accessed August 1, 2022) and annotating the protein domains using the protein domain database Pfam (http://pfam.xfam.org/; accessed August 1, 2022) via HMMSCAN in the HMMER package (http://hmmer.org/; accessed August 1, 2022).

2.4. Analysis of the public RNA-Seq data to search for novel candidate genes involved in immune responses

Articles containing RNA-Seq data related to immune response in *A. mellifera* were searched in PubMed (https://pubmed.ncbi.nlm.nih.gov/; accessed August 1, 2022) to retrieve immune-related RNA-Seq data for *A. mellifera* in public databases. The articles were curated and SRA files related to the articles were downloaded and converted to FASTQ files. The FASTQ files were quality-controlled and trimmed via Trim Galore! v. 0.6.7 (https://github.com/FelixKrueger/TrimGalore). Transcripts per million (TPM) values for the reference transcriptome based on the selected RNA-Seq data were calculated with kallisto v. 0.46.0 (https://github.com/pachterlab/kallisto; accessed August 1, 2022) [24].

To search for genes involved in immune response, the transcriptome scores for each experiment were calculated as shown in Equation (1) to avoid splitting by zero.

$$\text{Score} = \{(\text{average TPM values for pathogen-challenged samples}) + 1\} / \{(\text{average TPM values for control samples}) + 1\} \quad (1)$$

Transcripts with scores > 2 or < 0.5 per experiment were listed. Novel candidate genes were searched by comparing lists with Calculate and plotting Venn diagrams (https://bioinformatics.psb.ugent.be/webtools/Venn/; accessed August 1, 2022). For the gene set enrichment analyses, the transcripts were converted to human gene symbol annotations and used as the input for Metascape v. 3.5 (https://metascape.org/; accessed August 1, 2022)[25].

3. Results

3.1. Construction of reference transcriptome of A. mellifera using public RNA-Seq data

We curated all RNA-Seq data for *A. mellifera* in the GEO database to construct reference transcriptome sequences. From 50 available projects (~1,300 runs as of April 2021), we selected the SRA data for 35 projects (1,008 runs; Supplemental Table 1) after manual curation excluding duplicates, irregular conditions, and old types of NGS instruments. The curated RNA-Seq data were mapped to the genome data [3]. The reference transcriptome sequences were constructed using the mapped data and the reference *A. mellifera* gene model data. The transcriptome sequence contained 149,685 transcripts. There were 28,324 mRNAs in the gene model previously published in NCBI (GCF_003254395.2_Amel_HAv3.1_genomic.gtf) and designated RGM (reference gene model) (Figure 2A and Supplemental Data 1). There were 12,105 and 16,735 gene loci at the reference transcriptome and genome levels, respectively (Figure 2B and Supplemental Data 2). Thus, numerous gene isoforms predicted in RGM were identified in the reference transcriptome. Transcript IDs starting with "MSTRG" were newly identified in the present study.

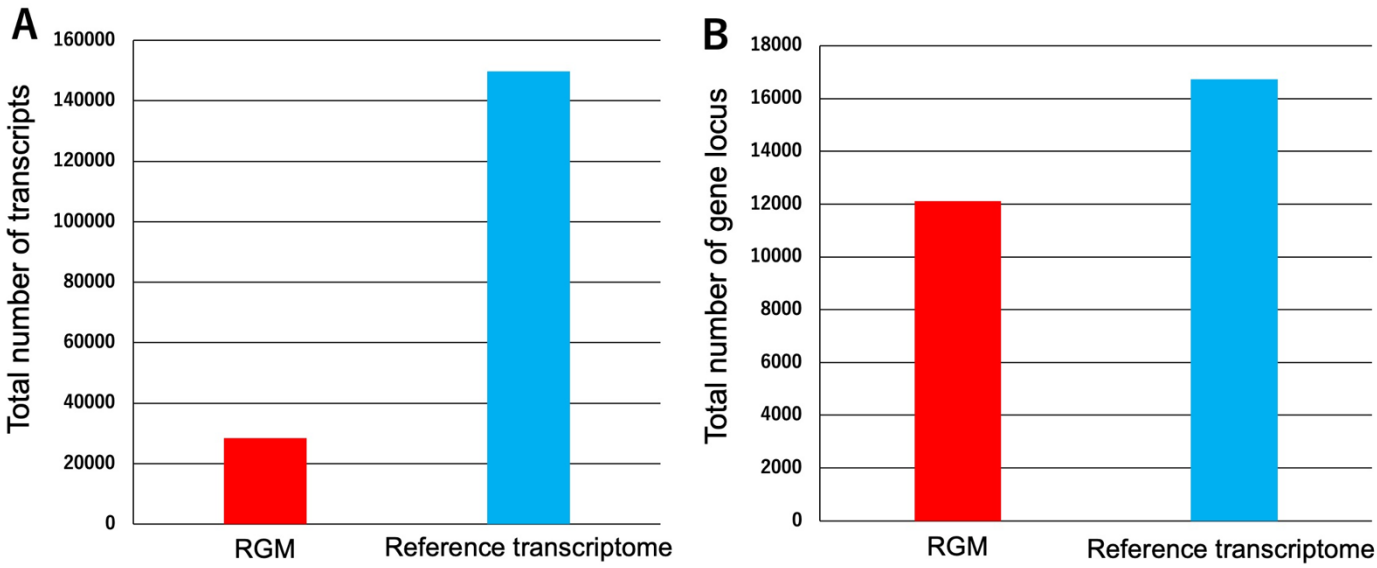

Figure 2. Comparisons of the total numbers of transcripts and gene loci in the reference gene model (RGM) and the reference transcriptome data. (A) Comparisons of the total numbers of transcripts and both datasets. (B) Comparisons of the total numbers of gene loci in both datasets via the gffcompare program.

We used the reference transcriptome sequences to predict the amino acid sequence with TransDecoder. We predicted 194,174 amino acid sequences from 116,863 transcript sequences (Supplemental Data 3). We evaluated the predicted data and the amino acid sequences of RGM via blastp analyses (Supplemental Data 4). The comparison revealed that 154,466 of the reference transcriptome amino acid sequences hit the RGM sequences whereas 23,454/23,471 of the RGM amino acid sequences hit the reference transcriptome sequences. We used Fanflow4Insects to conduct a functional annotation of the reference transcriptome using the amino acid sequences (Supplemental Data 5) [18]. Fanflow4Insects disclosed that ~62–47% of the amino acid sequences were functionally annotated via the *Homo sapiens, Mus musculus, and Caenorhabditis elegans* protein sequence datasets while ~62–50% were functionally annotated through the *Drosophila melanogaster, Bombyx mori, Bombus terrestis*, and *Nasonia vitripennis* protein sequence datasets (Table 1). We annotated 65.60% and 55.59% of the sequences with the UniGene and Pfam datasets, respectively.

Table 1. Number of annotated amino acid sequences.

Protein sequence dataset	*H. sapiens*	*M. musculus*	*C. elegans*	*D. melanogaster*	*B. mori*	*B. terrestris*	*N. vitripennis*	UniGene	Pfam
No. annotated predicted amino acid sequences (total 194,174)	121,027	115,617	91,859	97,544	121,119	113,306	110,591	127,386	107,936
% annotated sequences	62.33%	59.54%	47.30%	50.24%	62.38%	58.35%	56.95%	65.60%	55.59%

3.2. Meta-analysis of reference transcriptome and published RNA-Seq data to detect new candidate genes (transcripts) involved in the immune response

To determine the utility of the reference transcriptome data, we performed a meta-analysis to search new candidate genes (transcripts) using the constructed and public RNA-Seq data related to the immune response. To retrieve the RNA-Seq data, we searched articles in PubMed using the words "apis mellifera immune rna seq". We obtained 17 hits and curated them and their RNA-Seq data. We evaluated the article contents and the completeness and accessibility of the RNA-Seq dataset. After curation, we prepared the RNA-Seq data from five articles for use in the transcriptome expression analysis (Table 2). We prepared sequence data from the five bioprojects in the five articles. We calculated the TPM values of each RNA-Seq dataset after raw sequence data trimming and quality control. The TPM values of each experiment listed in Table 2 appear in Supplemental Data 6. We calculated the transcript scores for each experiment sought to identify those that were upregulated or downregulated in response to pathogen infection (Supplemental Data 7). Those that were upregulated and had transcript IDs with scores > 2 were listed in each experiment (Supplemental Data 8). No single upregulated transcript was common to all experiments. Thus, we searched for upregulated transcripts in the experiments related to bacterial infection (AFB, B_BN, P_PN, M_MN, and C_CN) or viral infection (CC_CV, RC_RV, SBV_DWV, and DWV). Figures 3A and 3B show that 20 and 10 transcripts were upregulated in the viral- and bacterial infection-related experiments, respectively. The emergent transcript IDs in each category are listed in Supplemental Data 9. We also searched downregulated transcripts with scores < 0.5. No single downregulated transcript was common to all experiments (Supplemental Data 9). Nine and three transcripts were downregulated in all viral infection- and bacterial infection-related experiments, respectively (Figures 3C and 3D; Supplemental Data 9).

Table 2. Public RNA-Seq data used in expression calculations.

Bioproject	**Sample**	**Treatments**	**Reference**	**Sequence Read Archive accession ID (sample (abbreviation)**	**Experimental abbreviation)**
PRJNA52851	4th instar honeybee larvae	*Paenibacillus larvae* orally inoculated and treated larvae were incubated 72 h.	[26]	SRR068395 (infected), SRR068396 (Control).	AFB
PRJEB6511	Newly hatched larvae	Sacbrood virus (SBV)-deformed wing virus (DWV) mixure orally inoculated and treated larvae were incubated	[27]	ERR528750 (control_1_D1), ERR528751 (control_2_D2), ERR528753 (DWV_SBV_1_C5), ERR528754 (DWV_SBV_2_C6),	SBV_DWV

Bioproject	Sample	Treatments	Reference	Sequence Read Archive accession ID (sample (abbreviation)	Experimental abbreviation)
		9 d. Control was PBS.			
PRJNA445764	Abdomens of worker adults	1. Diet, Pollen (P), Bee-Pro (B), MagaBee (M) or control (carbohydrate diet) (C) fed 7 d. 2. *Nosema apis* and sucrose provided. 3. Each diet fed 7 d.	[28]	SRR6901841 (B2), SRR6901842 (B1), SRR6901848 (B3), SRR6901856 (BN1), SRR6901857 (BN2), SRR6901858 (BN3) SRR6901843 (P1), SRR6901844 (P2), SRR6901849 (P3), SRR6901859 (PN1), SRR6901860 (PN2), SRR6901862 (PN3) SRR6901845 (M1), SRR6901846 (M2), SRR6901847 (M3), SRR6901853 (MN1) SRR6901854 (MN2) SRR6901855 (MN3) SRR6901850 (C1) SRR6901851 (C2) SRR6901852 (C3) SRR6901861 (CN1) SRR6901863 (CN2) SRR6901864 (CN3)	B_BN (control vs *N. apis*-inoculated samples reared with Bee-Pro) P_PN (control vs *N. apis*-inoculated samples reared with pollen) M_MN (control vs *N. apis*-inoculated samples reared with MagaBee) C_CN (control vs *N. apis*-inoculated samples reared with carbohydrate diet)
PRJNA498919	Whole bodies of adult bees	Samples reared with chestnut or rockrose (less nutritious), fed control or inoculated with Israeli acute paralysis virus (IAPV) containing 30% sucrose solution for 36 h	[29]	SRR8121077, SRR8121078 (CC_1), SRR8121079, SRR8121080 (CC_2), SRR8121081, SRR8121082 (CC_3), SRR8121083, SRR8121084 (CC_4), SRR8121085, SRR8121086 (CC_5), SRR8121087, SRR8121088 (CC_6), SRR8121101, SRR8121102 (CV_1), SRR8121103, SRR8121104 (CV_2), SRR8121105, SRR8121106 (CV_3), SRR8121107,	CC_CV (control vs IAPV-inoculated samples reared with chestnut)*

Bioproject	Sample	Treatments	Reference	Sequence Read Archive accession ID (sample (abbreviation)	Experimental abbreviation)
				SRR8121108 (CV_4), SRR8121109, SRR8121110 (CV_5), SRR8121111, SRR8121112 (CV_6)	
				SRR8121089, SRR8121090 (RC_1), SRR8121091, SRR8121092 (RC_2), SRR8121093, SRR8121094 (RC_3), SRR8121095, SRR8121096 (RC_4), SRR8121097, SRR8121098 (RC_5), SRR8121099, SRR8121100 (RC_6), SRR8121113, SRR8121114 (RV_1), SRR8121115, SRR8121116 (RV_2), SRR8121117, SRR8121118 (RV_3), SRR8121119, SRR8121120 (RV_4), SRR8121121, SRR8121122 (RV_5), SRR8121123, SRR8121124 (RV_6)	RC_RV (control vs IAPV-inoculated samples reared with rockrose)* *Two run files were tied to a single sample and merged into one file. TPMs were calculated by using merged data.
RJNA669279	Larvae	DWV containing PBS (or OBS as control) were inoculated 12 h in larvae reared for 3 d from 1-d larvae. Treated larvae were incubated 4 d.	[30]	SRR12830831 (control_1), SRR14424854 (control_2), SRR14424852 (control_3), SRR12830832 (DWV_1), SRR12830855 (DWV_2), SRR12830853 (DWV_3)	DWV (control vs DWV-infected)

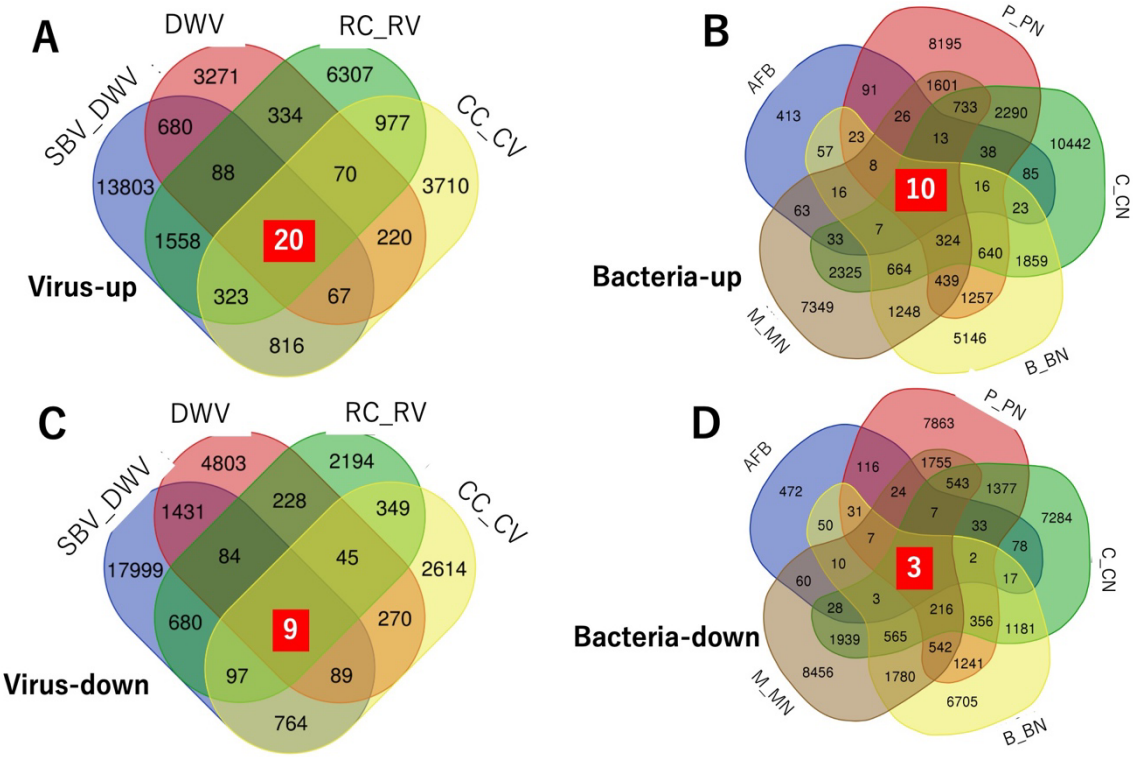

Figure 3. Transcripts upregulated or downregulated in response to viral or bacterial infection. Venn diagrams showing numbers of upregulated transcripts in experiments (Table 1) on viral infection (A) or bacterial infection (B) and numbers of downregulated transcripts in experiments on viral infection (C) or bacterial infection (D). White numbers in red squares indicate numbers of upregulated or downregulated transcripts in all experiments in (A)–(D).

We used the annotation data to examine the transcripts upregulated or downregulated in response to viral or bacterial infection (Supplemental Data 10). We used Metascape to perform an enrichment analysis on the 20 transcripts upregulated in response to viral infection. The inputs were annotations of human gene symbols converted by the annotation data. In the gene ontology (GO), the transcripts related to cellular protein catabolic process (GO:0044257) or protein ubiquitination (GO:00165567) were included in the 20 transcripts (Supplemental Figure 1). However, *P* values for both ontologies were not low. The transcripts upregulated in response to bacterial or viral infection did not include immune-related IMD, Toll, or JAK-STAT pathway or autophagy (Supplemental Data 10) [31,32]. However, aldehyde dehydrogenase 3 (XM_026442860.1) and negative regulator of RNA polymerase III (MSTRG.11647.15) were included among the transcripts upregulated in response to bacterial infection. By contrast, *autophagy related protein 3 (ATG3)* (MSTRG.1676.15) was included among the transcripts downregulated in response to viral infection while urea transporter transcript (MSTRG.12160.40) was downregulated in response to bacterial infection.

4. Discussion

The present study was a meta-analysis intended to provide reference transcriptome data for the honeybee *Apis mellifera*. We constructed the *A. mellifera* reference transcriptome sequence data using RNA-Seq data curated from the public database and reference chromosome-level genome sequence and their gene set data [3]. Fanflow4Insects was used for functional transcriptome annotation [18] integrating model species such as humans, *M. musculus*, *C. elegans*, and several insect species such as *B. mori*. We then performed another meta-analysis using reference transcriptome data constructed and public RNA-Seq data of immune-challenged *A. mellifera* to search novel candidate genes related to the immune response. We identified several candidate transcripts.

The reference transcriptome sequence consisted of 149,685 transcripts belonging to 16,735 loci. Whereas the RGM had 28,324 transcripts and 12,105 loci, the reference transcriptome had far greater numbers of both. Large numbers of RNA-Seq data from the public database were used as hint data to construct the reference transcriptome sequence. The public RNA-Seq data were derived from neurons, midguts, hypopharyngeal gland mushroom bodies, and other tissue samples at multiple developmental stages such as the embryo and the pupa (Supplemental Table 1). The abundant RNA-Seq data might contain the transcripts or isoforms expressed in the various tissues and developmental stages and possessing important biological functions. In mosquitoes, Down syndrome cell adhesion molecule (Dscam) mediates phagocytosis and is a pattern recognition receptor. It has many isoforms and is divergently associated with isoform-specific immune responses to various pathogens [33]. Thus, reference transcriptome data can reveal biological reactions related to isoform-specific functions. Moreover, the reference transcriptome includes loci previously unidentified in RGMs. We assumed that the locus transcripts were expressed at very low levels and at only a few developmental stages. The foregoing information suggests that transcriptome analyses using the reference transcriptome can accurately indicate entire expression states and disclose transcripts with isoform- and tissue-specific functions.

Functional transcriptome sequence annotations are critical in the biological interpretation of analytical results. For the functional annotation in the present study, 194,174 amino acid sequences from 116,863/149,685 transcripts were predicted because multiple amino acid sequences were predicted from a single transcript. Blastp against the RGM protein dataset demonstrated that the predicted amino acid sequences of the reference transcriptome covered nearly all the RGM protein data comprising 23,471 sequences. There were also 9,935 coding genes in the Refseq data (Amel_HAv3,1: NCBI Apis mellifera Annotation Release 104). The foregoing findings suggest that the reference transcriptome data covered nearly all the amino acid sequences for the genes and transcripts in the previous version and also included novel heretofore unidentified genes and transcripts. Fanflow4Insects [18] was used to search transcriptome orthologs among the model and insects species. There were hits between ~60% of the amino acid sequences of the transcriptome sequences and the human and *M. musculus* gene sets. There were also hits between 50% of them and the *D. melanogaster* gene set. A possible explanation is that the human and *M. musculus* gene sets were more sophisticated and better annotated than those of the other species. As more knowledge regarding these species is gained in the future, their annotations will improve. The functions of several amino acids were revealed but these data were not "functionally annotated" or incorporated into the public database. Several predicted amino acid sequences belonged to a single transcript. Usually, one matured mRNA encodes one amino acid. Therefore, some amino acid sequences from one transcript sequence may be incorrectly predicted. Hence such sequences cannot be functionally annotated, and the total percentage of hits might be reduced. The annotation results may be used in transcriptome enrichment analyses of GO and Kyoto Encyclopedia of Genes and Genomes (KEGG) annotations by converting the reference transcriptome ID to the gene symbol for the model species in the annotation file. Our amino acid sequence and annotation data may facilitate biological interpretations of the results of the transcriptome analyses.

We used reference transcriptome and RNA-Seq data from the public database to perform a meta-analysis and detect novel candidate genes (transcripts) related to the immune responses to viruses or bacteria [26–30]. Pathogens infecting honeybees are economically serious problems, bringing about mass deaths of honeybees. One of the resolutions to the problems is the enhancement of immune reactions of honeybees through gene modifications such as genome editing. Thus, searching for novel candidate genes related to the immune responses and understanding the immune system in honeybees can lead to the first step of the resolution, and we searched for the novel genes. We calculated the expression levels of the transcriptomes and the RNA-Seq data as well as the changes in transcriptome expression in response to viral or bacterial infection. We

applied the scores for the published immunological experiments to search the downregulated and upregulated transcripts common to them all. However, we found none. We then searched the upregulated or downregulated transcripts common to four viral or five bacterial experiments. There were 20 upregulated and nine downregulated transcripts among the viral experiments and 10 upregulated and three downregulated transcripts among the bacterial experiments. Hence, there were very few transcripts and none of them was functionally annotated as immune-related. Possible explanations for these discrepancies were the facts that the experiments widely varied with respect to incubation time, *A. mellifera* developmental stage, inoculation procedure and difference of beekeeping styles between laboratories. In addition, *A. mellifera* samples are usually prepared from different colonies with diverse genetic backgrounds and are therefore genetically diverse (meaning genetic variations in individual samples), which can affect gene expressions. A potential reason for the latter discrepancy is that *A. mellifera* must be bred using worker bees and not under strictly controlled conditions. The TPM values of the identified transcripts widely varied. Nevertheless, the candidate transcripts could participate in the immune response as they emerged under highly diverse conditions. *ATG3* was included among the common transcripts downregulated in response to viral infection and may be implicated in ubiquitin-like conjugation systems that prolong the autophagosomal membrane phase. The autophagosomal membrane phase is one of three autophagy processing stages. *ATG3* functions in *D. melanogaster* [34]. Autophagy suppresses viral replication. When autophagy-related genes were silenced in *D. melanogaster* challenged with vesicular stomatitis virus (VSV), the insects presented with higher mortality rates than the control. We assumed that in *A. mellifera*, *ATG3* was downregulated in response to viral infection to avoid the host immune response and autophagy. When Toll-7 detected VSV in *Drosophila,* autophagy activation did not depend on canonical immune signaling pathways such as IMD or Toll [34]. This finding was consistent with that of our meta-analysis, which detected no immune-related transcripts among those commonly upregulated or downregulated in response to microbial infections. The foregoing results suggest that autophagy was implicated in the viral immune responses in *A. mellifera* and autophagy regulation regardless of the canonical immune pathways are conserved in both *D. melanogaster and A. mellifera*. In future work, meta-analysis using small numbers of RNA-Seq data in fewer studies, more relevant RNA-Seq data, or RNA-Seq data using pure ligands such as peptidoglycan or lipopolysaccharide could reveal new immune-related genes.

Half of the immune-related-candidate genes (transcripts) were not included in RGM (IDs of them start "MSTRG".). The reason for it can be that the reference transcriptome includes many transcripts which are not included in RGM. As described above, the reference transcriptome data were constructed by using many RNA-Seq data as hint data derived from multiple tissues, developmental stages, different castes, and the bees under many conditions (e.g., infecting bacteria), leading to identifications of the new transcripts. In addition, we performed annotations of the transcriptome sequences by using gene sets of model species (e.g., human and mouse) and several model insect species, resulting in ~60 % transcriptome sequences annotation [18]. Eventually, *ATG3* transcript described in the previous paragraph was newly identified in the reference transcriptome, and the "*ATG3*" annotation was derived from human gene sets. Considering them, transcriptome analysis using either public RNA-Seq data or newly sequenced RNA-Seq data and the reference transcriptome data could lead to new biological insights and interpretations of honeybees.

In the present study, we constructed reference transcriptome data and showed that they could readily be applied towards a meta-analysis. The results of the latter showed that transcriptome expression widely varied in response to microbial infection in *A. mellifera*. In future work, RNA-Seq data of *A. mellifera* under strictly controlled conditions and at all developmental stages will be need. In this manner, the dynamics of transcriptome expression in *A. mellifera* may be elucidated. Our transcriptome data could also be used in these RNA-Seq data analyses or RNA-Seq analyses of conditioned samples

such as pathogen-infected worker bees. In addition, comparative transcriptome analysis between castes using the reference transcriptome data and RNA-Seq of each caste could reveal the molecular mechanism of the differences in the biological features among castes, such as differences in immune responses and morphological, behavioral, and physiological differences. Since the transcriptome constructed by RNA-Seq of diverse honeybees derived from multiple castes can include these isoforms, our reference transcriptome could contribute to the identifications of caste-specific isoforms involved in the differences among castes. These investigations will help clarify the molecular biology of *A. mellifera*.

Author Contributions: Conceptualization, K.Y. and H.B.; methodology, K.Y. and H.B.; software, K.Y, T.W., and H.B.; resource, K.Y. and H.B.; validation, K.Y., T.W., and H.B; formal analysis, K.Y., T.W., and H.B; investigation, K.Y. and T.W.; data curation, K.Y., T.W., and H.B; writing—original draft preparation, K.Y.; writing—review and editing, K.Y., T.W. and H.B.; visualization, K.Y.; supervision, H.B.; project administration, H.B.; funding acquisition, H.B. All authors have read and agreed to the final draft and the published version of the manuscript.

Acknowledgments: This work was based on the results obtained for Project JPNP19001 subsidized by the New Energy and Industrial Technology Development Organization (NEDO).

Funding: This research was funded by Project JPNP19001 subsidized by the New Energy and Industrial Technology Development Organization (NEDO) and the Center of Innovation for Bio-Digital Transformation (BioDX) which is an open innovation platform for industry-academia co-creation (COI-NEXT) of the Japan Science and Technology Agency (JST, COI-NEXT, JPMJPF2010) to H.B.

Conflicts of Interest: The authors have no conflict of interest to declare.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: All supplemental data are available in figshare (DOI:10.6084/m9.figshare.c.6047651).

Supplemental Table 1: RNA-Seq data used in this study for hint data (DOI: 10.6084/m9.figshare.20072387).

Supplemental Data 1: Reference transcriptome sequences data in gtf and fasta file formats (DOI: 10.6084/m9.figshare.20264418).

Supplemental Data 2: Output gffcompare file using gtf files of reference transcriptome and gene model (DOI: 10.6084/m9.figshare.20279856).

Supplemental Data 3: Predicted amino acid data for *Apis mellifera* transcriptome in gtf and fasta file formats (DOI: 10.6084/m9.figshare.20272698).

Supplemental Data 4: Comparisons of predicted amino acid data of reference transcriptome and amino acid sequences of gene models performed by blastp (DOI: 10.6084/m9.figshare.20291931).

Supplemental Data 5: Annotations of reference *Apis mellifera* transcriptome by Fanflow4Insects [18] (DOI: 10.6084/m9.figshare.20272749).

Supplemental Data 6: TPM values of transcripts in each experiment. {experimental name}_tpm.tsv contains TPM values of all transcripts in samples of experiments (DOI: 10.6084/m9.figshare.20485272).

Supplemental Data 7: Scores of transcripts in each experiment. {experimental name}_score.tsv contains scores for all transcripts in each experiment (DOI: 10.6084/m9.figshare.20485305).

Supplemental Data 8: Transcript IDs upregulated or downregulated in response to pathogen infection. Transcripts IDs with scores > 2 and < 0.5 were listed in {experimental name}_up.txt and {experimental name}_down.txt, respectively (DOI: 10.6084/m9.figshare.20485329).

Supplemental Data 9: List of common upregulated or downregulated transcripts common to all, viral infection-related, or bacterial infection-related experiments (DOI: 10.6084/m9.figshare.20485359).

Supplemental Data 10: Annotation lists of transcripts upregulated or downregulated in response to viral or bacterial infections. Annotation lists of transcripts upregulated or downregulated in response to either viral or bacterial infections. Transcripts not shown in list were not annotated or translated to amino acids (DOI: 10.6084/m9.figshare.20486553).

Supplemental Figure 1: Enrichment analysis by Metascape using 20 transcripts upregulated in response to viral infections. Human gene symbols of 20 transcripts were used as input data and enrichment analysis was performed in Metascape. Enrichment analysis by Metascape using 20 transcripts downregulated in response to viral infections. $-\log_{10}$ (P) represents $-\log_{10}$ (p-value). For example, $-\log_{10}$ (p) = 2 represents p-value = 10^{-2} (DOI: 10.6084/m9.figshare.20486508).

References

1. Winston, M. *The Biology of the Honey Bee*; Harvard University Press, 1991.
2. Weinstock, G.M.; Robinson, G.E.; Gibbs, R.A.; Weinstock, G.M.; Weinstock, G.M.; Robinson, G.E.; Worley, K.C.; Evans, J.D.; Maleszka, R.; Robertson, H.M. Insights into social insects from the genome of the honeybee Apis mellifera. *Nature*. **2006**, *443*, 931–949. DOI:10.1038/nature05260.
3. Wallberg, A.; Bunikis, I.; Pettersson, O.V.; Mosbech, M.B.; Childers, A.K.; Evans, J.D.; Mikheyev, A.S.; Robertson, H.M.; Robinson, G.E.; Webster, M.T. A hybrid de novo genome assembly of the honeybee, Apis mellifera, with chromosome-length scaffolds. *BMC Genomics*. **2019**, *20*, 275. DOI:10.1186/s12864-019-5642-0.
4. Schlüns, H.; Crozier, R.H. Relish regulates expression of antimicrobial peptide genes in the honeybee, Apis mellifera, shown by RNA interference. *Insect Mol. Biol.* **2007**, *16*, 753–759. DOI:10.1111/j.1365-2583.2007.00768.x.
5. Bono, H. All of gene expression (AOE): An integrated index for public gene expression databases. *PLOS ONE*. **2020**, *15*, e0227076. DOI:10.1371/journal.pone.0227076.
6. Fent, K.; Schmid, M.; Hettich, T.; Schmid, S. The neonicotinoid Thiacloprid causes transcriptional alteration of genes associated with mitochondria at environmental concentrations in honey bees. *Environ. Pollut.* **2020**, *266*, 115297. DOI:10.1016/j.envpol.2020.115297.
7. Decio, P.; Ustaoglu, P.; Derecka, K.; Hardy, I.C.W.; Roat, T.C.; Malaspina, O.; Mongan, N.; Stöger, R.; Soller, M. Thiamethoxam exposure deregulates short ORF gene expression in the honey bee and compromises immune response to bacteria *[Sci. rep.]. Sci. Rep.* **2021**, *11*, 1489. DOI:10.1038/s41598-020-80620-7.
8. Soares, M.P.M.; Pinheiro, D.G.; de Paula Freitas, F.C.; Simões, Z.L.P.; Bitondi, M.M.G. Transcriptome dynamics during metamorphosis of imaginal discs into wings and thoracic dorsum in Apis mellifera castes. *BMC Genomics*. **2021**, *22*, 756. DOI:10.1186/s12864-021-08040-z.
9. Yi, Y.; Liu, Y.B.; Barron, A.B.; Zeng, Z.J. Transcriptomic, morphological, and developmental comparison of adult honey bee queens (Apis mellifera) reared from eggs or worker larvae of differing ages. *J. Econ. Entomol.* **2020**, *113*, 2581–2587. DOI:10.1093/jee/toaa188.
10. Kawamoto, M.; Jouraku, A.; Toyoda, A.; Yokoi, K.; Minakuchi, Y.; Katsuma, S.; Fujiyama, A.; Kiuchi, T.; Yamamoto, K.; Shimada, T. High-quality genome assembly of the silkworm, Bombyx mori. *Insect Biochem. Mol. Biol.* **2019**, *107*, 53–62. DOI:10.1016/j.ibmb.2019.02.002.
11. Yokoi, K.; Tsubota, T.; Jouraku, A.; Sezutsu, H.; Bono, H. Reference transcriptome data in silkworm Bombyx mori. *Insects*. **2021**, *12*, 519. DOI:10.3390/insects12060519.
12. Li, Q.; Li, M.; Zhu, M.; Zhong, J.; Wen, L.; Zhang, J.; Zhang, R.; Gao, Q.; Yu, X.Q.; Lu, Y. Genome-wide identification and comparative analysis of cry toxin receptor families in 7 insect species with a focus on Spodoptera litura. *Insect Sci.* **2022**, *29*, 783–800. DOI:10.1111/1744-7917.12961.
13. Masuoka, Y.; Cao, W.; Jouraku, A.; Sakai, H.; Sezutsu, H.; Yokoi, K. Co-expression network and time-course expression analyses to identify silk protein regulatory factors in Bombyx mori. *Insects*. **2022**, *13*, 131. DOI:10.3390/insects13020131.
14. Ono, Y.; Bono, H. Multi-Omic meta-analysis of transcriptomes and the Bibliome uncovers novel hypoxia-inducible genes. *Biomedicines*. **2021**, *9*, 582. DOI:10.3390/biomedicines9050582.
15. Suzuki, T.; Ono, Y.; Bono, H. Comparison of oxidative and hypoxic stress responsive genes from meta-analysis of public transcriptomes. *Biomedicines*. **2021**, *9*, 1830. DOI:10.3390/biomedicines9121830.
16. Yokoi, K.; Kimura, K.; Bono, H. Revealing Landscapes of Transposable Elements in Apis species by Meta-analysis **2022**, *Insects*. 13(8), 698; DOI: 10.3390/insects13080698 .
17. Doublet, V.; Poeschl, Y.; Gogol-Döring, A.; Alaux, C.; Annoscia, D.; Aurori, C.; Barribeau, S.M.; Bedoya-Reina, O.C.; Brown, M.J.F.; Bull, J.C.; et al. Unity in Defence: Honeybee Workers Exhibit Conserved Molecular Responses to Diverse Pathogens. *BMC Genomics* **2017**, 18, 207, DOI:10.1186/s12864-017-3597-6.
18. Bono, H.; Sakamoto, T.; Kasukawa, T.; Tabunoki, H. Systematic functional annotation workflow for insects. *Insects*. **2022**, *13*, 586. DOI:10.3390/insects13070586.

19. Pertea, M.; Kim, D.; Pertea, G.M.; Leek, J.T.; Salzberg, S.L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and ballgown. *Nat. Protoc.* **2016**, *11*, 1650–1667. DOI:10.1038/nprot.2016.095.
20. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinform.* **2009**, *25*, 2078–2079. DOI:10.1093/bioinformatics/btp352.
21. Pertea, G.; Pertea, M. *GFF Utilities: GffRead and GffCompare 2020*.
22. Haas, B.J.; Papanicolaou, A.; Yassour, M.; Grabherr, M.; Blood, P.D.; Bowden, J.; Couger, M.B.; Eccles, D.; Li, B.; Lieber, M.; et al. De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. *Nat. Protoc.* **2013**, *8*, 1494–1512. DOI:10.1038/nprot.2013.084.
23. Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J.. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. DOI:10.1093/nar/25.17.3389.
24. Bray, N.L.; Pimentel, H.; Melsted, P.; Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **2016**, *34*, 525–527. DOI:10.1038/nbt.3519.
25. Zhou, Y.; Zhou, B.; Pache, L.; Chang, M.; Khodabakhshi, A.H.; Tanaseichuk, O.; Benner, C.; Chanda, S.K. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* **2019**, *10*, 1523. DOI:10.1038/s41467-019-09234-6.
26. Cornman, R.S.; Lopez, D.; Evans, J.D. Transcriptional response of honey bee larvae infected with the bacterial pathogen Paenibacillus larvae. *PLOS ONE*. **2013**, *8*, e65424. DOI:10.1371/journal.pone.0065424.
27. Ryabov, E.V.; Fannon, J.M.; Moore, J.D.; Wood, G.R.; Evans, D.J. The Iflaviruses Sacbrood virus and deformed wing virus evoke different transcriptional responses in the honeybee which may facilitate their horizontal or vertical transmission. *PeerJ*. **2016**, *4*, e1591. DOI:10.7717/peerj.1591.
28. Azzouz-Olden, F.; Hunt, A.; DeGrandi-Hoffman, G. Transcriptional response of honey bee (Apis mellifera) to differential nutritional status and nosema infection. *BMC Genomics*. **2018**, *19*, 628. DOI:10.1186/s12864-018-5007-0.
29. Rutter, L.; Carrillo-Tripp, J.; Bonning, B.C.; Cook, D.; Toth, A.L.; Dolezal, A.G. Transcriptomic responses to diet quality and viral infection in Apis mellifera. *BMC Genomics*. **2019**, *20*, 412. DOI:10.1186/s12864-019-5767-1.
30. Chang, Z.T.; Huang, Y.F.; Chen, Y.W.; Yen, M.R.; Hsu, P.Y.; Chen, T.H.; Li, Y.H.; Chiu, K.P.; Nai, Y.S. Transcriptome-level assessment of the impact of deformed wing virus on honey bee larvae *[Sci. rep.]. Sci. Rep.* **2021**, *11*, 15028. DOI:10.1038/s41598-021-94641-3.
31. Lemaitre, B.; Hoffmann, J. The host defense of Drosophila melanogaster. *Annu. Rev. Immunol.* **2007**, *25*, 697–743. DOI:10.1146/annurev.immunol.25.022106.141615.
32. Kuo, C.J.; Hansen, M.; Troemel, E. Autophagy and innate immunity: Insights from invertebrate model organisms. *Autophagy*. **2018**, *14*, 233–242. DOI:10.1080/15548627.2017.1389824.
33. Dong, Y.; Taylor, H.E.; Dimopoulos, G. AgDscam, a hypervariable immunoglobulin domain-containing receptor of the Anopheles gambiae innate immune system. *PLOS Biol.* **2006**, *4*, e229. DOI:10.1371/journal.pbio.0040229.
34. Moy, R.H.; Cherry, S. Antimicrobial autophagy: A conserved innate immune response in Drosophila. *J. Innate Immun.* **2013**, *5*, 444–455. DOI:10.1159/000350326.