*insects*

*Article*

# Novel RNA Viruses from the Transcriptome of Pheromone Glands in the Pink Bollworm Moth, *Pectinophora gossypiella*

**Xiaoyi Dou [1], Sijun Liu [1,\*], Victoria Soroker [2], Ally Harari [2] and Russell Jurenka [1]**

[1]   Department of Entomology, Iowa State University, Ames, IA 50011, USA; xydou@uga.edu (X.D.);
      rjurenka@iastate.edu (R.J.)
[2]   The Volcani Center, Institute of Plant Protection, ARO, Bet-Dagan 50250, Israel;
      sorokerv@volcani.agri.gov.il (V.S.); aharari@volcani.agri.gov.il (A.H.)
\*    Correspondence: sjliu42@gmail.com

**Simple Summary:** The pink bollworm, *Pectinophora gossypiella* (Lepidoptera: Gelechiidae), is a major pest of cotton. In this study, we analyzed the mRNA from pheromone glands of two populations in Israel. We found several virus sequences that were the same in these populations. We identified these viruses based on high-throughput sequencing data and analysis of the assembled transcripts. Through analysis of the sequences, we identified several unique viral sequences representing possible novel viral species. Two of the viral sequences were found in relatively high abundance in pheromone glands. One of the virus sequences was also found through analysis of previous transcriptome sequencing data from the midgut of pink bollworm larvae. This is the first report of these unique viral sequences found in the pink bollworm, and these viruses could be developed to help control this pest around the world, but more research is needed to determine their utility as biological control agents.

**Abstract:** In this study, we analyzed the transcriptome obtained from the pheromone gland isolated from two Israeli populations of the pink bollworm *Pectinophora gossypiella* to identify viral sequences. The lab population and the field samples carried the same viral sequences. We discovered four novel viruses: two positive-sense single-stranded RNA viruses, Pectinophora gossypiella virus 1 (PecgV1, a virus of *Iflaviridae*) and Pectinophora gossypiella virus 4 (PecgV4, unclassified), and two negative-sense single-stranded RNA viruses, Pectinophora gossypiella virus 2 (PecgV2, a virus of *Phasmaviridae*) and Pectinophora gossypiella virus 3 (PecgV3, a virus of *Phenuiviridae*). In addition, sequences derived from two negative-sense single-stranded RNA viruses that belong to *Mononegavirales* were found in the data. Analysis of previous transcriptome sequencing data derived from the midgut of pink bollworm larvae of a USA population only identified PecgV1, but no other viruses. High viral sequence coverages of PecgV1 and PecgV4 were observed in both field and lab populations. This is the first report of viral sequences discovered from the pink bollworm. Results from this investigation suggest that the pink bollworm harbors multiple viruses. Further investigation of the viral pathogens may help to develop novel pest management strategies for control of the pink bollworm.

**Keywords:** single-stranded RNA virus; iflavirus; pink bollworm; transcriptome

## 1. Introduction

The pink bollworm (PBW), *Pectinophora gossypiella* (Lepidoptera: Gelechiidae), is a worldwide pest that causes significant yield losses in cotton fields. The larvae of the pink bollworm feed inside flower buds and bolls, where they are well protected from contact to insecticides; hence, application of traditional insecticides is not efficient in controlling the pest. Current control measures for PBW largely depend on the use of transgenic cotton producing the *Bacillus thuringiensis* (Bt) toxin, which effectively controls this pest [1]. The use of Bt cotton showed a long-term suppression of pink bollworm in

a ten-year study in Arizona, USA [1]. However, PBW populations resistant to Bt toxin Cry1Ac have been observed in the field [2]. In addition, artificial sex pheromones used in mating disruption have been applied in many areas. They successfully controlled PBW populations in the cotton-growing areas of the USA, Egypt and Israel for decades [3]. Recently, resistant populations to sex pheromones have been found in cotton fields of Israel, which compromises the use of this control measure [4]. The use of viral pathogens to manage PBW was also investigated. As with other lepidopteran insects, the larvae of PBW are susceptible to nuclear polyhedrosis viruses (baculoviruses). For instance, *Autographa californica nuclear phlyhedrosis virus* (AcNPV) was tested in bait formulation against PBW larvae, which resulted in significantly decreased numbers of larvae and thus boll damage [5,6].

Viral pathogens of PBW were previously recorded. The infectivity, symptomatology, histopathology and transmission of AcNPV in pink bollworm larvae were investigated [7]. In addition to the DNA virus, a small RNA virus (picoRNA-like virus) infecting pink bollworm larvae and adults and causing disease symptoms and even death of the infected larvae was isolated from cotton fields of Egypt. This virus infected midgut cells and could be vertically transmitted from infected adults to offspring [8]. However, genomic sequences of the RNA virus have not been reported.

Assembly of high-throughput sequencing data provides a powerful methodology for discovery of viral sequences from insects [9]. Analysis of the assembled transcripts to identify novel viruses is now a common tool for determination of viral sequences and for construction of full-length viral genomes of RNA viruses. Furthermore, viral sequence analysis from the transcriptome often identified multiple viruses that co-infected host insects without causing disease symptoms [10]. Due to the difference in the sex pheromone ratio between laboratory and field populations, and since there are no significantly different expressed genes involved in sex pheromone biosynthesis [4], we attempted to investigate transcriptome variations in viruses in the pheromone glands. Here, we observed significant numbers of contigs that hit viral sequences by BLAST annotations. Through analysis of the putative viral sequences, we identified putative viral sequences representing possible novel viral species, and we herein describe their unique features.

## 2. Materials and Methods

### 2.1. Insect Collection and Pheromone Gland Extraction

Two different populations of pink bollworm were collected in Israel. One was from a laboratory colony (Lab) maintained at the Volcani Institute, Bet-Dagan, Israel, and the other one was collected from a cotton field (Field) near Ein Shemer, Israel, where mating disruption failed [4]. The laboratory population was maintained in the lab for many years but with the periodical addition of field-caught insects from Israel that had never been exposed to the PBW pheromone. The field insects originated from a cotton field in Israel and were reared in the lab to increase the population. Pheromone glands along with ovipositors from adult females of the two populations were removed, immediately placed in RNAlater (Invitrogen), frozen at −80 °C and then shipped to the lab in the United States.

### 2.2. RNA Isolation and Illumina Sequencing

Methods for RNA isolation, sequencing library preparation and Illumina sequencing were described previously [4]. Briefly, total RNA from 30 pheromone glands was isolated using TRIzol reagent (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's protocol. A 2100 Bioanalyzer (Agilent Technologies) was used to check the quantity of RNA. After removing rRNA, the mRNA library was prepared and sequenced on an Illumina HiSeq 3000 platform by the DNA facility at Iowa State University, Ames, IA, USA. The stranded total RNA-Seq libraries were sequenced with 150 paired ends and the mRNA library with 100 single reads. All sequencing reads were submitted to the SRA of NCBI under the accession number "SRP140160". The total number of reads generated from the mRNA library was $3.38 \times 10^8$ and $3.35 \times 10^8$ in lab and field populations, while the

stranded total RNA-Seq libraries were $8.14 \times 10^7$, $5.55 \times 10^7$ and $4.88 \times 10^7$ in the lab population and $9.91 \times 10^7$, $6.38 \times 10^7$ and $4.86 \times 10^7$ in the field population.

*2.3. De Novo Assembly of Short Reads and Gene Annotation*

The quality of all raw reads was checked using FastQC (Babraham Bioinformatics, Cambridge, UK). The de novo assembly was carried out with the short reads assembling program Trinity [11]. To search potential RNA sequences derived from viruses, the assembled contigs were scanned for candidate sequences by BLASTx (E-value < 0.01) in a search against a local virus sequence database, which was organized from NCBI viral sequences (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus? (1 July 2015)). Contigs that hit viral sequences were selected and further annotated by using the BLASTx program against the NCBI non-redundant (*nr*) protein databases. Details of the viral sequence search and determination were previously described [12]. Viral sequence mapping was performed by using a Perl script, which only mapped perfectly matched reads to the viral sequences [12].

*2.4. Sequence and Phylogenetic Analysis*

The conserved and functional domains of the predicted viral proteins were identified by the Conserved Domain Database (CDD) in NCBI (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi (accessed on 14 June 2021) and InterProtScan (https://www.ebi.ac.uk/interpro/search/sequence-search (accessed on 14 June 2021). The RNA-dependent RNA polymerase domain of picorna-like viruses was determined by the online NCBI conserved domain search engine (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi? (accessed on 14 June 2021) with an E-value of <0.001. The identified RdRP sequences were extracted and used for phylogenetic analysis. Multiple sequence alignments of the core motifs were conducted using CLC genomics Workbench 9.5. For the phylogenetic analysis, the sequences were aligned using the ClustalW method. MEGA 7.0 [13] was used for tree construction using the neighbor-joining method with 1000 bootstrap replicates, the Poisson model and pairwise gap deletion options.

## 3. Results

*3.1. Identification of Sequences Derived from Putative Viruses*

The initial screen of potential virus-related sequences by BLASTx against the local viral databases resulted in the selection of ~1% of contigs for further analysis. The selected contigs were further annotated by a BLAST search of the NCBI nr protein databases. The contigs of putative viral origins with a nucleotide length longer than 500 and an amino acid length longer than 100 were selected. In total, there were 19 transcript hits to viruses in the lab population and 21 hits (length > 500 nt) to viruses in the field population (Table 1). Out of the viral sequences found, there were 9 and 11 transcripts of dsDNA viruses in the lab and field populations, respectively, and 8 were homologous to ssRNA viruses in both populations, with 2 of them having positive-sense and 8 negative-sense RNA genomes (Tables S1 and S2). Among those sequences, two viral contigs derived from (+ss)RNA viruses encoded entire viral genes and contained 5′- and 3′-untranslated regions (UTR); hence, these sequences were near full-length virial genomes. There were eight contigs with (−ss)RNA viral sequences that also encoded complete viral genes and hit two (−ss)RNA viruses. Those contigs were also likely derived from viral genomes. In addition, two putative viral transcripts encoding protein sequences were found that hit the RdRp of Formica fusca virus 1 (Contig 1 in lab population and Contig 18 in field population); however, the encoded viral protein appeared truncated. Hence, these two contigs were likely derived from viral sequences integrated in the host genomes. Based on the viral genome structure and BLAST search results, these eight viral RNAs were likely derived from two (+ss) and six (−ss) RNA viruses. We then named the two (+ss)RNA viruses as Pectinophora gossypiella virus1 (PecgV1) and Pectinophora gossypiella virus 4 (PecgV4), and two (−ss)RNA viruses, Pectinophora gossypiella virus 2 (PecgV2) and Pectinophora gossypiella virus 3 (PecgV3). The sequences of these eight ssRNAs were

submitted to GenBank under "MN164617 to MN164623 and MZ361082 and MZ361083". The other two (−ss)RNA viruses belong to *Mononegavirales*.

**Table 1.** Number of viral RNA fragments found in the transcriptome of pheromone gland in two populations.

| | Virus | dsDNA Virus | ssRNA Virus | |
|---|---|---|---|---|
| | | | Positive-Sense | Negative-Sense |
| Lab | 19 | 9 | 2 | 8 |
| Field | 21 | 11 | 2 | 8 |

Amino acid alignment of the six ssRNA viruses between the lab and field populations showed 99% to 100% sequence identity (Table 2). To assess the relative amount of viral RNA in the field and lab populations, we mapped the sequencing reads to the genomes of the RNA viruses. Remarkably, the two (+) ssRNA viruses accumulated very high viral sequences. Totals of 3.34% and 1.09% of the reads were derived from PecgV1 in field and lab populations, respectively, which are equivalent to 39,250 and 4052 folds of sequence coverages. Sequencing reads derived from PecgV4 also contributed to 1.1% and 2.5% of the total reads. The base coverage of PecgV1 was up to 8475-fold for the field sample and 5511-fold for the lab sample. Accumulations of the viral RNA from PecgV2 and PecgV3 were much lower comparing to the two positive RNA viruses. However, the viral sequence coverages of these two viruses were variable between 50- and 710-fold (Table 2). These results indicate that PecgV1 and PecgV4 were highly accumulated in the pheromone gland tissues.

**Table 2.** Comparison of ssRNA viruses between two populations of the PBW.

| ssRNA Viruses | AA Length | Identity (Field vs. Lab) | Sequence Coverage (Field/Lab) | % of Reads (Field/Lab) |
|---|---|---|---|---|
| PecgV1 | 2948 | 99% | 39,250.6/4052 | 3.34/1.09 |
| PecgV2-L | 2093 | 99% | 87.4/56 | 0/0 |
| PecgV2-M | 724 | 99% | 49.3/168.9 | 0/0 |
| PecgV2-S | 364 | 99% | 286.7/487.3 | 0/0 |
| PecgV3-S | 277 | 100% | 106.1/50 | 0/0 |
| PecgV3-M | 1558 | 99% | 170.8/103.2 | 0/0 |
| PecgV3-L | 2502 | 99% | 709.9/632.4 | 0.05/0.14 |
| PecgV4 | 2846 | 99% | 8474.6/5510.7 | 1.1/2.25 |

In the NCBI/SRA database, we found transcriptomic sequencing data from larval PBW isolated from a USA population (SRA: SRP062867) [14]. We assembled the transcriptome reads, and only one transcript of 10,031 nt was annotated as a putative viral sequence. Sequence annotation showed the transcript to share 99% sequence identity to PecgV1, indicating that this transcript was derived from PecgV1. Hence, PecgV1 is likely distributed widely in PBW populations, while PecgV2, PecgV3 and PecgV4 were only found in the Israeli populations.

### 3.2. Analysis of Positive-Sense Single-Stranded RNA Viruses

BLASTp analysis of the amino acid sequences indicated that PecgV1 was a member of *Iflaviridae*. Although protein annotation of PecgV4 showed a close relationship to iflaviruses, its genome orientation is different from that of an iflavirus. For example, the structural and non-structural proteins are arranged in the opposite position in PecgV4 (Figure 1). This picorna-like virus was observed in recently discovered unclassified insect picorna-like viruses, Diaphorina citir picorna-like virus (DcPLV) (KT698837.1) [15], Riptortus pedestris virus-2 (RpV2) (MN078225.1) and Hubei picorna-like virus 33 (HPLV33) (NC_033210.1) [16]. The replication-related L-fragment of PecgV2 and PecgV3 hit the families of *Phasmaviridae* and *Phenuiviridae*, respectively.
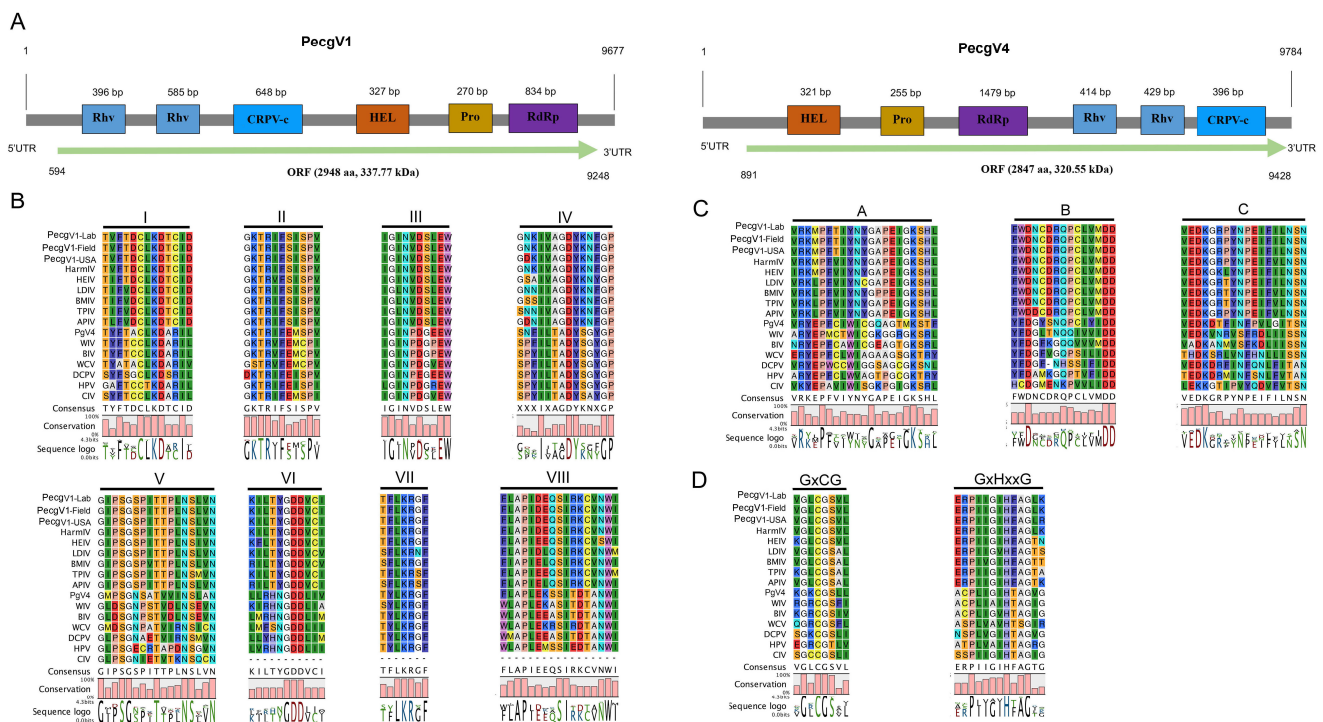
**Figure 1.** Genome organization and amino acid alignment of PecgV1 and PecgV4. (**A**): CDD and InterProScan analysis of genome structure of PecgV1 and PecgV4. Rhv: Picornavirus capsid protein domain_like; CRPV_c: CRPV capsid protein like; Hel: RNA helicase; Pro: protease; RdRp: RNA-dependent RNA polymerase; UTR: untranslated region; ORF: open reading frame. (**B**): RdRp amino acid alignment. (**C**): Helicase amino acid alignment. (**D**): Protease amino acid alignment.

The PecgV1 polyprotein suggested that PecgV1 is highly homologous to an iflavirus (Helicoverpa armigera iflavirus) isolated from Helicoverpa armigera with 78% identity (Tables S1 and S2), which represents a possible new species in *Iflaviridae*. The genome of 9677 nt comprised a 593-nt 5′ UTR with a putative internal ribosome entry site (IRES), a single ORF polyprotein of 2948 aa with a predicted molecular weight of 337.77 kDa and a 429-nt 3′ UTR followed by a poly A tail (Figure 1A). PecgV4 is 9784 nt in length and also encoded a polyprotein of 2846 aa with a predicted molecular weight of 320.55 kDa. It is notable that PecgV4 has a different protein orientation comparing to iflaviruses in which its capsid protein domains were located in the C′-terminus (Figure 1C). The structural and non-structural protein domains of the PecgV4 polyprotein showed ~40% identity to the counterparts of the polyprotein of Lysiphlebus fabarum rna-virus type A, which is an iflavirus. The other top hits of the PecgV4 polyprotein were also from Iflaviridae.

Despite the differences in polyprotein structures of PecgV1 and PecgV4, the conserved protein domains identified from PecgV1 and PecgV4 were similar (Figure 1A), including the RNA helicase (pfam00910), protease (peptidase_C3G super family) (cl13774) and RdRp (cd01699) domains in the non-structural polyprotein half, and two Rhv-like (cd00205) and a CrPV-coat capsid (cl07393) domains in the structural polyprotein half. All of these domains are common in iflaviruses. Further analysis identified eight picornaviral conserved motifs of RdRP domains (Figure 1B), three conserved motifs from RNA helicase domains (Figure 1C) and two conserved protease motifs (Figure 1D) from PecgV1 and PecgV4. These motifs were previously described by Koonin et al. [17]. Sequence comparison of the motifs among PecgV1 and several iflaviruses from Helicoverpa armigera, Bombyx mori, Lymantria dispar, Antheraea pernyi and Thaumetopoea pityocampa showed highly conserved amino acid compositions in those motifs (Figure 1). In contrast, however, the amino acid motifs of PecgV4 were more divergent compared to those of PecgV1 (Figure 1).

Phylogenetic analysis of RdRP domains of PecgV1 and PecgV4 with known iflaviruses, dicistroviruses and other unclassified picorna-like viruses revealed that PecgV1 is closely

related to the viruses isolated from *H. armigera* and *B. mori*. The three iflaviruses together with three lepitopteran iflaviruses form a subclade (Figure 2). Although PecgV4 is structurally similar to DcPLV, HPLV33 and RpV2, the RdRP of PecgV4 is grouped with several unclassified insect viruses with iflaviral genome structures (Figure 2); phylogenetically, PecgV4 is more closely related to iflaviruses than to the viruses which have the same genome structures as PecgV4.
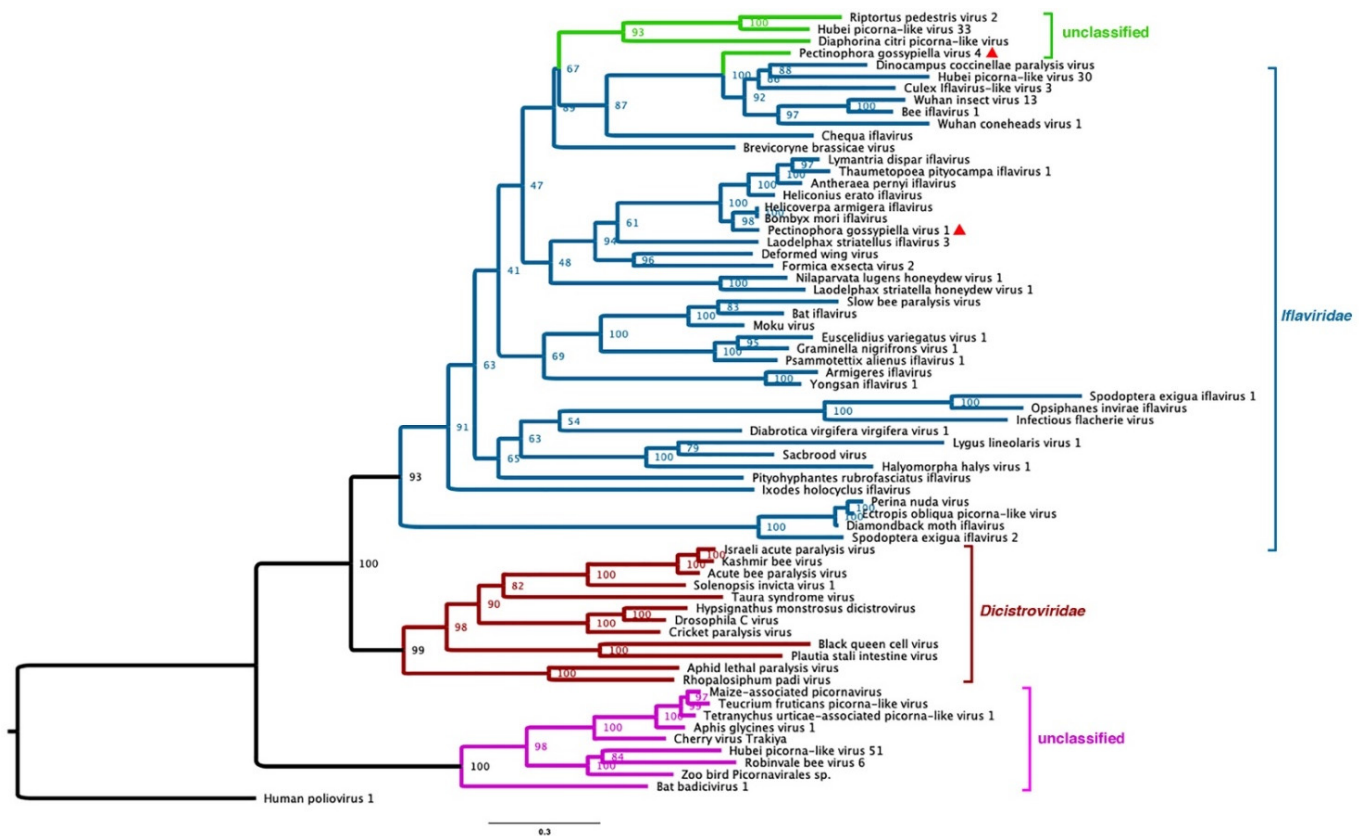


**Figure 2.** Phylogenetic tree of PecgV1 and PecgV4 (red triangle) with other iflaviruses. RdRP domain sequences (cd) were used. The accession numbers for sequences are listed in File S1. The amino acid sequences were aligned using ClastalW, and then the tree was constructed using Mega 7.0 with neighbor-joining method and evaluated with 1000 bootstrap replicates.

### 3.3. Negative-Sense Single-Stranded RNA Viruses

The initial annotation of PecgV2 and PecgV3 suggested that these two viruses were (-ss)RNA viruses belonging to *Bunyavirale*. The genomes of bunyaviruses consist of three subgenomic RNA segments: large, medium and small (L, M and S). The L segment encodes RdRP, and M and S encode glycoproteins and nucleocapsid proteins, respectively [18]. For PecgV2, the large segment was 6419 nt in length, which contained a complete CDS of RdRp. The RdRp of PecgV2 is 2094 aa with a predicted molecular weight of 241.89 kDa (Figure 3A). A Bunya_RdRp domain (pfam0419, interval 618–1231aa) was determined by CDD analysis with an E-value of $8.09 \times 10^{-7}$. BLAST analysis of the RdRP protein of PecgV2 showed a high amino acid identity (62%) to that of Seattle prectang virus (SPV), which is a member of the family *Phasmaviridae* and was isolated from the green pug (*Pasiphila retangulata*), a lepidopteran. The M segment of PecgV2 was 2446 nt in length. The top BLASTp hit was 42.47% identity to the glycoprotein of Seattle prectang virus. The conserved domain prediction showed that there is a phlebovirus glycoprotein G2 domain in the middle region (interval: 268–525, E-value: $1.64 \times 10^{-3}$). The S segment of PecgV3 was 1701 nt, encoding a potential nucleoprotein with 40.46 kDa. The BLASTp result showed 56.74% identity to the nucleocapsid of Seattle prectang virus; however, CDD analysis did not find any conserved domains.
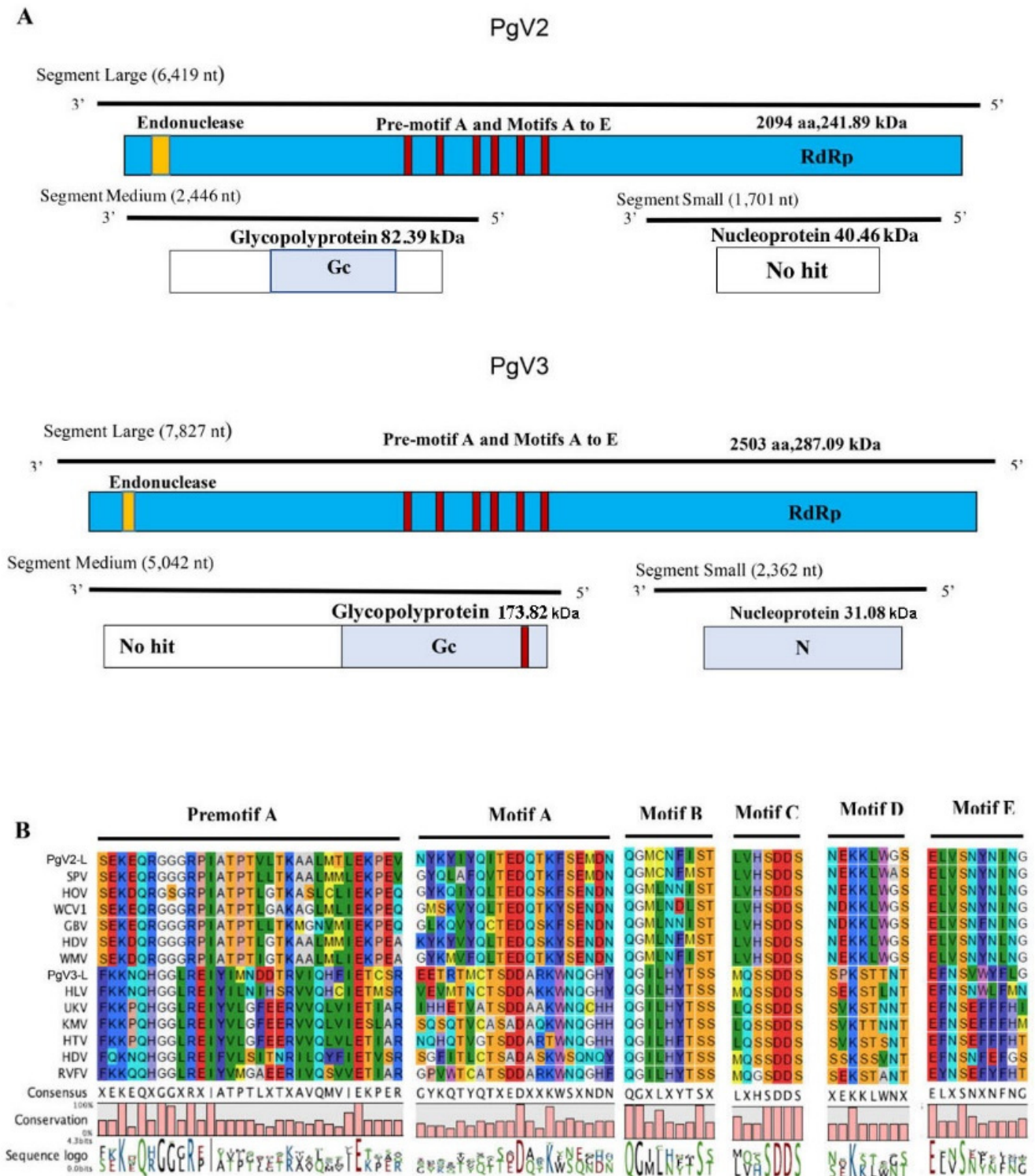
**Figure 3.** Genome organization and amino acid alignment of PecgV2 and PecgV3. (**A**): CDD and InterProScan analysis of genome structure of PecgV2 (top) and PecgV3 (below). (**B**): Amino acid alignment between conserved RdRp premotif A and motifs A–E of PecgV2 and PecgV3, and selected (−ss)RNA viruses. The accession numbers of sequences for amino acid alignment are listed in File S1; red bars represent the conserved motifs in the genome.

The genome of PecgV3, with its three segments, was recovered from the sequencing reads. The three encoded viral proteins hit Hubei lepidoptera virus 1 (HLV1) of *Bunyavirale* by the BLASTp search. The L segment was 7827 nt in length and encoded a putative RdRp of 2503 aa with a predicted molecular weight of 287.09 kDa (Figure 3A). The L

segment of PecgV2 and PecgV3 only shared 12% amino acid identity, indicating they are not closely related. The conserved Bunya_RdRp L protein domain was predicted to be between 809 and 1485aa (E-value $3.68 \times 10^{-74}$). The alignment of the RdRp from PecgV2-L and PecgV3-L and other (−ss)RNA viruses indicates that the amino acid sequences of their RdRps consist of the pre-motif A and motifs A through E, which are highly conserved in negative-sense RNA viral polymerases and common to all other *Bunyavirales* characterized to date [19] (Figure 3B). We found three basic residues (K, R and R/K) in promotif A and a glutamic acid (E) downstream of premotif A, which are conserved in bunyaviral RdRps [20,21]. SDD sequences in motif C were also found in both viruses. The other motifs were slightly different between these two viruses. The M segment of PecgV3 was 5042 nt in length and encoded a surface glycoprotein of 1558 aa with a mass of 173.82 kDa. BLASTp results show that the M segment had 27% identity to the putative glycoprotein of HLV1. The conserved domain predicted by CDD showed a single transmembrane region close to the C-terminus of the glycoprotein (interval: 1052–1520, E-value: $1.14 \times 10^{-15}$). The online programs TMHMM and TMPred showed there are two transmembrane domains in the middle region and one transmembrane domain in the C-terminus. This result is expected as multiple internal transmembrane domains are conserved in *Phenuiviridae*. The S segment of PecgV3 is 2362 nt in length and encoded for a unique protein of 276 aa with a predicted mass of 31.08 kDa. Protein annotation showed that it was closely associated with the N protein of members of the genus *Phasivirus*. CDD analysis showed there is a *Tenuivirus/Phlebovirus* nucleocapsid protein domain (interval: 23–245, E-value: $3.98 \times 10^{-6}$).

Phylogenetic trees were constructed based on the amino acid sequences of the RdRp segment (Figure 4). The accession numbers of the other viruses are shown in File S1. PecgV2 RdRp was clustered together with the family *Phasmaviridae* in the order *Bunyavirales*. PecgV3 was clustered together with HLV1, a member in the family *Phenuiviridae*. All the data show that PecgV2 and PecgV3 are different viruses and both belong to the order *Bunyavirales*, but in different families.

### 3.4. Endogenous Viral Elements (EVE) Derived from RNA Viruses

In mammals, retroviral envelope genes play an important role in cell–cell fusion during placentation, and they might have anatomic and immunological functions [22]. The two contigs (lab contig 1 and field contig 18; lab contig 17 and field contig 19) that hit viral genes and were annotated as a putative EVE were both derived from mononegaviruses, a viral group consisting of one segment of the (−)ssRNA genome. The initial annotation of lab contig 1/field contig 18, a 2325-nt RNA fragment, suggested that the encoded CDS was likely derived from a nucleocapsid gene (N) of a virus close to Spodoptera frguiperda rhabdovirus. Further analysis by BLASTp against nr databases revealed that the encoded protein is a homolog of uncharacterized proteins from *Ostrinia furnacalis*, *Papilio* spp., *Aphantopus hyperantus* and other insects. Lab contig 17/field contig 19 was 4423 nt in length, encoding two CDSs. They both hit RdRP of Formica exsecta virus 4 (FeV4) or Formica exsecta virus 1 (FeV1), members of *Monoegavirales*. A full-length RdRP of FeV4 or FeV1 is 1877aa; hence, the two CDS of the contigs are actually truncated sequences. In addition, no sequences of other viral genes related to the RdRPs were assembled from the reads, and these results suggest that the RdRP hit contigs were putatively derived from the EVE of the host.
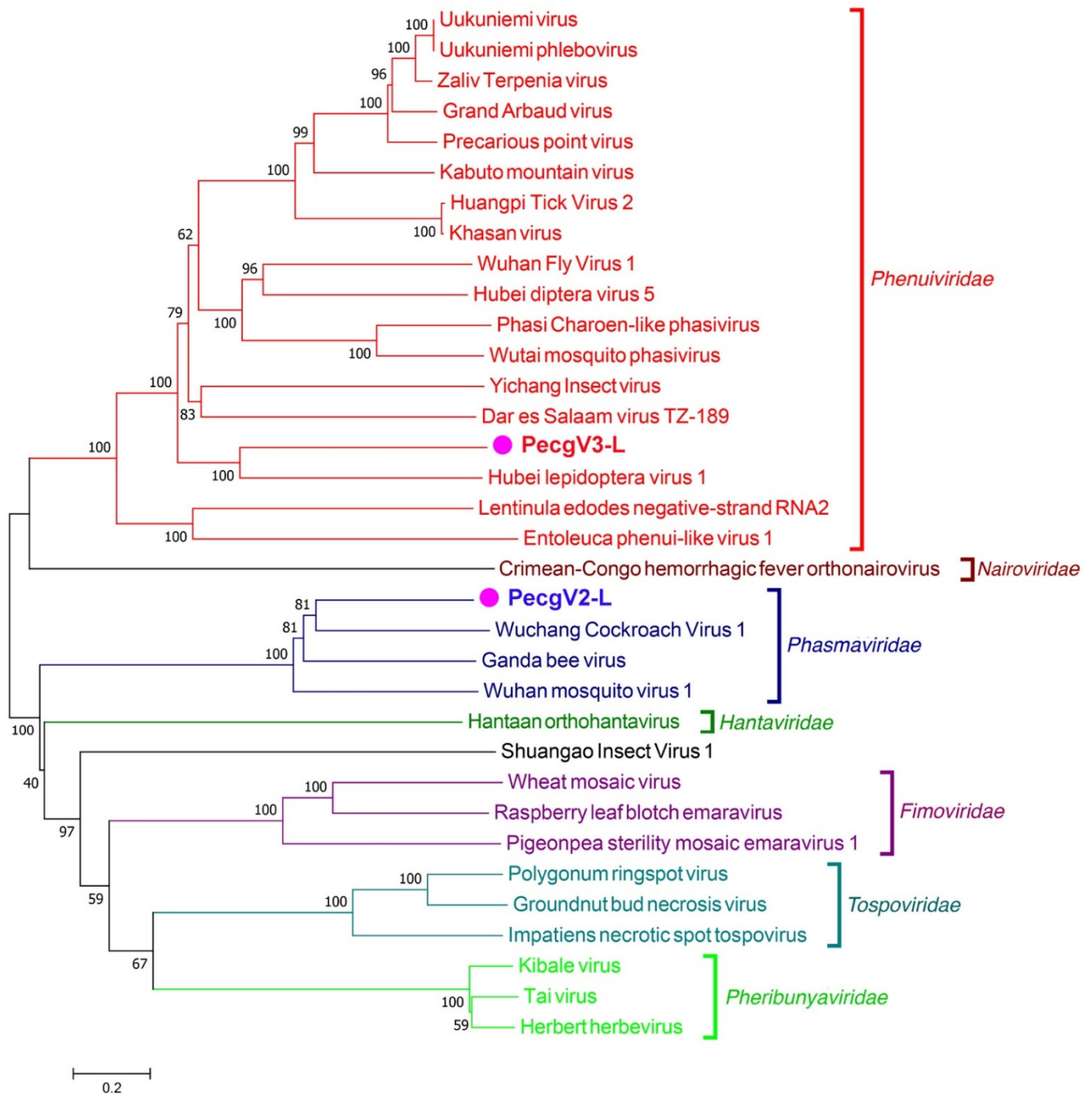
**Figure 4.** Phylogenetic trees of PecgV2 and PecgV3 with other (−ss)RNA with RdRp sequences. The accession numbers for sequences are shown in File S1. The amino acid sequences were aligned using ClastalW, and then the tree was constructed using Mega 7.0 with neighbor-joining method and evaluated with 1000 bootstrap replicates.

## 4. Discussion

New viruses are difficult to discover with the traditional viral detection methods because they require isolation of viruses followed by whole genome sequencing. However, high-throughput next-generation sequencing technologies and bioinformatics have been applied to virus discovery in various organisms, including humans [23,24], arthropods [25–27], plants [28,29] and fungi [30–32]. Transcriptome sequencing and deep sequencing of viral small RNA analysis have also been used for the discovery of viruses in plant and insect hosts [33,34].

The pink bollworm, as a pest species, quickly spread around the world as it was first reported on infested cotton in Egypt in 1913 [35] and in 1917 in the USA [36]. In the present

study, through transcriptome analysis, we discovered novel RNA viral sequences likely derived from viruses infecting the pink bollworm and derived from mononegaviruses (EVE). The viral sequences were not different between the lab and field populations from Israel. The four viruses share 99% to 100% sequence identity in the comparison of lab and field populations. Higher viral sequence coverages from the sequencing reads indicate that they are consistently expressed in the pink bollworm. However, fewer viral sequences were recovered from the American larval sample analyzed by Tassone et al. [14]. Only PecgV1 was observed in both the Israeli and American samples. The transcriptome of the Israeli population was from the pheromone gland with ovipositors, while the USA population was from the midgut of larvae. We assume that PecgV1 is continuously infecting PBW in different development stages, at least in adults and larvae, and it is passed to the next generation vertically. The fewer viral sequences in the American samples might be due to the lower sequence coverage in the larva RNA sample, different insect stages or various other potential factors. It is interesting that extremely high accumulations of the viral RNA derived from PecgV1 and PecgV4 were found in the pheromone glands, suggesting that these viruses replicated at high levels in adult pink bollworms. However, no visible disease symptoms were observed in the adults. Hence, the potential biological functions of the viruses to the hosts remain to be investigated as well as the tissue distribution within the host.

To date, four genomic types of picornaviruses isolated from insects have been identified. These viruses share similar conserved domains: the RdRp (cd0169) and Rh-V domains and other motifs. Members of *Iflaviridae* and *Dicistroviridae* could cause serious disease symptoms to the host insect. For instance, *Aphid lethal paralysis virus* (ALPV), a dicistrovirus, could significantly reduce aphid populations [37,38]. Dicistroviruses have two ORFs: ORF1 encodes non-structural polyproteins, while ORF2 expresses structural proteins. The other commonly found picorna-like viruses are iflaviruses. The genome of iflaviruses only expresses a single ORF, encoding the non-structural proteins (NSPs); the structural proteins (SPs) are encoded in one ORF [39]. The NSPs are located in the N-terminal of the protein sequence, while the C-terminal contains SPs, and these are one of the genomic characteristics of iflaviruses. Recently, two new types of unclassified insect-infecting picorna-like viruses were discovered by analyzing assembled transcripts form transcriptome sequencing data. Aphis glycines virus 1 (Accession KM013260.2) represents a new group of dicistro-like viruses, in which the genome encodes two ORFs, but is different from dicistroviruses in that the SPs are located in the N-terminal and the NSPs are in the C-terminal half of the protein sequence. PecgV4, DcPLV, HPLV33 and RpV2 are members of a group of picorna-like viruses, whose genomes encode a single ORF with NSP in the N-terminal half and SP in the C-terminal half of the protein sequence. Phylogenetic analysis of RdRp domains indicated that PecgV1 and the viruses with PecgV4-like genomes are close to iflaviruses, but distant from dicistroviruses (Figure. 2), suggesting that the two groups of picorna-like viruses which encode only a single ORF might have evolved from the same ancestors, although the NSP and SP are in opposite locations for the two groups of viruses, while the two groups of viruses in which the NSP and SP are independently encoded were closely related in viral evolution. The phylogenetic analysis based on the RdRp domains showed that the unclassified dicistro-like viruses appeared earlier in the phylogeny, suggesting that the other picorna-like viral groups might have evolved from the newly discovered dicistro-like viruses (Figure 2).

PecgV2 and PecgV3 were defined as (-ss)RNA viruses. Based on the amino acid alignment and phylogenetic analysis, it is clear that the sequences of PecgV2 and PecgV3 were derived from different viruses. PecgV2 is a virus in the family *Phasmaviridae.* Viruses of this family are known to be isolated from insect hosts, while PecgV3 relates to viruses in the family *Phenuiviridae*, which infect both vertebrates and invertebrates. Hence, the pink bollworm is likely the host of PecgV2 and PecgV3. It has been proposed that plant negative RNA viruses evolved via horizontal virus transfer from vertebrates and mediated by arthropod vectors. Therefore, it is reasonable that members of *Bunyavirales*, mostly

detected in plants and vertebrates, can be found in invertebrates, such as the pink bollworm in this study and other insects. The ancestors of plant and vertebrate negative RNA viruses are thought to have evolved from a natural reservoir which is represented by invertebrates via horizontal virus transfer, indicating that diverse invertebrates show similar viruses in the same ecological niche [40]. Investigation of the viruses in other insects would help to give a strong support of this view.

EVE sequences were commonly observed in insect genomes [41]. In addition to retroviral-related EVEs, EVEs derived from DNA and RNA viruses and (+) ssRNA, (-) ssRNA and double-stranded (ds)RNA viruses were observed in the same insect species. Most of the EVE sequences were disrupted in the host genomes [12]. Rhabdovirus-like EVE was previously observed in *S. frugiperda* cells [42]. Rhabdorivirus EVE derived from Spodoptera frugiperda rhabdovirus was also observed in RNA-seq of *S. frugiperda* larvae (Liu, unpublished results). We identified two transcripts likely derived from EVEs of unknown rhabdoviruses. The transcript encoding the N gene also hit other uncharacterized genes of several lepidopterans (data not shown), suggesting that the EVEs of mononegaviruses are common in these insects. Recovery of transcripts of the EVEs suggested that the EVEs were active. However, the coding regions of the contigs encoded two truncated RdRp sequences, indicating that the partial RdRp might not have biological functions. Some evidence has pointed to RNA-level functions for insect nonretrovial EVEs [43,44], but whether the EVE transcripts were translated and have functions in PBW needs to be further investigated.

In this study, we found the PBWs were infected by multiple RNA viruses. Picorna-like virus isolated from the pink bollworm population of Egypt was previously reported [8]. An Egyptian isolate of picorna-like virus caused severe pathogenetic impacts on the insects, specifically infection of the virus, resulting in death of the diseased larvae [8]. It is not clear whether the presently identified PecgV1 and PecgV4 were the same viruses as the viruses reported in 1995. Infection with picorna-like viruses does not always cause disease symptoms in the host insect. The dicistrovirus ALPV has been isolated from many insect species but was only observed causing diseases in the field population of cherry oat aphids (*Rhopasiphum padi*) [37,38]. However, it was asymptomatic in isolates from laboratory populations of pea aphids (*Acyrthosiphon pisum*) and *R. padi* [45]. Hence, it is possible that PecgV1 or PecgV4 may be an isolate of the previously observed picorna-like virus in Egypt. Our research provides the viral sequences that have potential for development of novel biological control agents for management of this worldwide devastating cotton insect pest.

## References

1. Carrière, Y.; Ellers-Kirk, C.; Sisterson, M.; Antilla, L.; Whitlow, M.; Dennehy, T.J. Long-term regional suppression of pink bollworm by *Bacillus thuringiensis* cotton. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 1519–1523. [CrossRef] [PubMed]
2. Tabashnik, B.E.; Liu, Y.B.; Dennehy, T.J.; Sims, M.A.; Sisterson, M.S.; Biggs, R.W.; Carrière, Y. Inheritance of resistance to Bt toxin Cry1Ac in a field-derived strain of pink bollworm (Lepidoptera: Gelechiidae). *J. Econ. Entomol.* **2002**, *95*, 1018–1026. [CrossRef] [PubMed]
3. Critchley, B.R.; Campion, D.G.; Mcveigh, L.J.; Hunter-Jones, P. Control of pink boll worm, *Pectinophora gossypiella* (Saunders) (Lepidoptera: Gelechiidae), in Egypt by mating disruption using an aerially applied microencapsulated pheromone formulation. *Bull. Entomol. Res.* **1983**, *73*, 289–299. [CrossRef]
4. Dou, X.; Liu, S.; Soroker, V.; Harari, A.; Jurenka, R. Pheromone gland transcriptome of the pink bollworm moth, *Pectinophora gossypiella*: Comparison between a laboratory and field population. *PLoS ONE* **2019**, *14*, e0220187. [CrossRef] [PubMed]
5. Bell, M.R.; Kanavel, R.F. Potential of bait formulations to increase effectiveness of nuclear polyhedrosis virus against the pink bollworm. *J. Econ. Entomol.* **1975**, *68*, 389–391. [CrossRef]
6. Bell, M.R.; Kanavel, R.F. Field tests of a nuclear polyhedrosis virus in a bait formulation for control of pink bollworms and *Heliothis* spp. in cotton in Arizona. *J. Econ. Entomol.* **1997**, *70*, 625–629. [CrossRef]
7. Vail, P.V.; Jay, D.L.; Hunter, D.K.; Staten, R.T. A nuclear polyhedrosis virus infective to the pink bollworm, *Pectinophora gossypiella*. *J. Invertebr. Pathol.* **1972**, *20*, 124–128. [CrossRef]
8. Monsarrat, A.; Abol-Ela, S.; Abdel-Hamid, I.; Fediere, G.; Kuhl, G.; EI Husseini, M.; Giannotti, J. A new RNA picorna-like virus in the cotton pink bollworm *Pectinophora gossypiella*. (LEP: Gelechiidae) in Egypt. *Entomophaga* **1995**, *40*, 47–54. [CrossRef]
9. Liu, S.; Vijayendran, D.; Bonning, B.C. Next generation sequencing technologies for insect virus discovery. *Viruses* **2011**, *3*, 1849–1869. [CrossRef]
10. Liu, S.; Chen, Y.; Bonning, B.C. RNA virus discovery in insects. *Curr. Opin. Insect Sci.* **2015**, *8*, 54–61. [CrossRef]
11. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.; et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **2011**, *29*, 644–652. [CrossRef]
12. Liu, S.; Coates, B.S.; Bonning, B.C. Endogenous viral elements integrated into the genome of the soybean aphid, *Aphis glycines*. *Insect Biochem. Mol. Biol.* **2020**, *123*, 103405. [CrossRef]
13. Kumar, S.; Strcher, G.; Tamura, K. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **2016**, *33*, 1870–1874. [CrossRef]
14. Tassone, E.E.; Zastrow-Hayes, G.; Mathis, J.; Nelson, M.E.; Wu, G.; Lindsey Flexner, J.; Carriere, Y.; Tabashnik, B.E.; Fabrick, J.A. Sequencing, de novo assembly and annotation of a pink bollworm larval midgut transcriptome. *GigaScience* **2016**, *5*, 28. [CrossRef]
15. Nouri, S.; Salem, N.; Nigg, J.C.; Falk, B.W. Diverse array of new viral sequences identified in worldwide population of the Asian citrus psyllid (*Diaphorina citri*) using viral metagenomics. *J. Virol.* **2016**, *90*, 2434–2445. [CrossRef] [PubMed]
16. Shi, M.; Lin, X.D.; Tian, J.H.; Chen, L.J.; Chen, X.; Li, C.; Qin, X.; Li, J.; Cao, J.; Eden, J.; et al. Redefining the invertebrate RNA virosphere. *Nature* **2016**, *540*, 539–543. [CrossRef] [PubMed]
17. Koonin, E.V.; Dolja, V.V.; Morris, T.J. Evolution and taxonomy of positive-strand RNA viruses: Implications of comparative analysis of amino acid sequences. *Crit. Rev. Biochem. Mol. Biol.* **2018**, *28*, 375–430. [CrossRef] [PubMed]
18. Amroun, A.; Priet, S.; de Lamballerie, X.; Querat, G. Bunyaviridae RdRps: Structure, motifs, and RNA synthesis machinery. *Crit. Rev. Microbiol.* **2017**, *43*, 753–778. [CrossRef] [PubMed]
19. Reguera, J.; Weber, F.; Cusack, S. Bunyaviridae RNA polymerases (L-protein) have an N-terminal, influenza-like endonuclease domain, essential for viral cap-dependent transcription. *PLoS Pathog.* **2010**, *6*, e1001101. [CrossRef]
20. Bruenn, J.A. A structural and primary sequence comparison of the viral RNA dependent RNA polymerases. *Nucleic Acids Res.* **2003**, *31*, 1821–1829. [CrossRef]
21. Elbeaino, T.; Diagiaro, M.; Alabdullah, A.; De Stradis, A.; Minafra, A.; Mielke, N.; Castellano, M.A.; Martelli, G.P. A multipartite single-stranded negative-sense RNA virus is the putative agent of Figure mosaic disease. *J. Gen. Virol.* **2009**, *90*, 1281–1288. [CrossRef]
22. Parrish, N.F.; Tomonaga, K. Endogenized viral sequences in mammals. *Curr. Opin. Microbiol.* **2016**, *31*, 176–183. [CrossRef]
23. Chiu, C.Y. Viral pathogen discovery. *Curr. Opin. Microbiol.* **2013**, *16*, 468–478. [CrossRef]
24. Lipkin, W.I.; Firth, C. Viral surveillance and discovery. *Curr. Opin. Virol.* **2013**, *3*, 199–204. [CrossRef]
25. Kobayashi, D.; Isawa, H.; Fujita, R.; Murota, K.; Itokawa, K.; Higa, Y.; Katayama, Y.; Sasaki, T.; Mizutani, T.; Iwanage, S.; et al. Isolation and characterization of a new iflavirus from *Armigeres* spp. mosquitoes in the Philippines. *J. Gen. Virol.* **2017**, *98*, 2876–2881. [CrossRef]
26. Smith, G.; Macias-Munoz, A.; Briscoe, A.D. Genome sequence of a novel iflavirus from mRNA sequencing of the butterfly *Heliconius erato*. *Genome Announc.* **2014**, *2*, e0039814. [CrossRef] [PubMed]
27. Sparks, M.E.; Gundersen-Rindal, D.E.; Harrison, R.L. Complete genome sequence of a novel iflavirus from the transcriptome of *Halyomorpha halys*, the brown marmorated stink bug. *Genome Announc.* **2013**, *1*, e00910-13. [CrossRef]
28. Barba, M.; Czosnek, H.; Hadidi, A. Historical perspective, development and application of next-generation sequencing in plant virology. *Viruses* **2014**, *6*, 106–136. [CrossRef] [PubMed]

29. Hany, U.; Adams, I.P.; Glover, R.; Bhat, A.I.; Boonham, N. The complete genome sequence of piper yellow mottle virus (PYMoV). *Arch. Virol.* **2014**, *159*, 385–388. [CrossRef]

30. Chiapello, M.; Rodríguez-Romero, J.; Ayllón, M.A.; Turinam, M. Analysis of the virome associated to grapevine downy mildew lesions reveals new mycovirus lineages. *Virus Evol.* **2020**, *6*, veaa058. [CrossRef]

31. Nerva, L.; Forgia, M.; Ciuffo, M.; Chitarra, W.; Chiapello, M.; Vallino, M.; Varese, G.C.; Turina, M. The mycovirome of a fungal collection from the sea cucumber *Holothuria polii*. *Virus Res.* **2019**, *273*, 197737. [CrossRef]

32. Sutela, S.; Forgia, M.; Vainio, E.J.; Chiapello, M.; Daghino, S.; Vallino, M.; Martino, E.; Girlanda, M.; Perotto, S.; Turina, M. The virome from a collection of endomycorrhizal fungi reveals new viral taxa with unprecedented genome organization. *Virus Evol.* **2020**, *6*, veaa076. [CrossRef]

33. Li, R.; Gao, S.; Hernandez, A.G.; Wechter, W.P.; Fei, Z.; Ling, K.S. Deep sequencing of small RNAs in tomato for virus and viroid identification and strain differentiation. *PLoS ONE* **2012**, *7*, e37127. [CrossRef]

34. Zografidis, A.; Nieuwerburgh, F.V.; Kolliopoulou, A.; Apostolou-Karampelis, K.; Head, S.R.; Defore, D.; Smagghe, G.; Swevers, L. Viral small-RNA analysis of *Bombyx mori* larval midgut during persistent and pathogenic cytoplasmic polyhedrosis virus infection. *J. Virol.* **2015**, *89*, 11473–11486. [CrossRef]

35. Ballou, H.A. The pink bollworm (*Gelechia gossypiella*) in Egypt. *J. Econ. Entomol.* **1918**, *11*, 236–245. [CrossRef]

36. Hunter, W.D. *The Pink Bollworm with Special Reference to the Steps Taken by the Department of Agriculture to Prevent its Establishment in the United States*; Bulletin 1397; U.S. Department of Agriculture: Washington, DC, USA, 1926; p. 12.

37. D'Arcy, C.J.; Burnett, P.A.; Hewings, A.D.; Goodman, R.M. Purification and characterization of a virus from the aphid *Rhopalosiphum padi*. *Virology* **1981**, *112*, 346–349. [CrossRef]

38. Williamson, C.; Rybicki, E.P.; Kasdorf, G.G.F.; Von Wechmar, M.B. Characterization of a new picorna-like virus isolated from aphids. *J. General Virol.* **1988**, *69*, 787–795. [CrossRef]

39. Hashimoto, Y.; Watanabe, A.; Kawase, S. In vitro translation of infectious acherie virus RNA in a wheat germ and a rabbit reticulocyte system. *Biochim. Biophys. Acta* **1984**, *781*, 76–80. [CrossRef]

40. Dolja, V.V.; Koonin, E.V. Metagenomics reshapes the conceptd of RNA virus evolution by revealing extensive horizontal virus transfer. *Virus Res.* **2018**, *244*, 36–52. [CrossRef]

41. Katzourakis, A.; Gifford, R.J. Endogenous viral elements in animal genomes. *PLoS Genet.* **2010**, *6*, e1001191. [CrossRef]

42. Geisler, C.; Jarvis, D. Rhabdovirus-like endogenous viral elements in the genome of *Spodoptera frugiperda* insect cells are actively transcribed: Implications for adventitious virus detection. *Biologicals* **2016**, *44*, 219–225. [CrossRef]

43. Ter Horst, A.; Nigg, J.C.; Dekker, F.M.; Falk, B.W. Endogenous viral elements are widespread in arthropod genomes and commonly give rise to PIWI-interacting RNAs. *J. Virol.* **2019**, *93*, e02124-18. [CrossRef]

44. Tassetto, M.; Kunitomi, M.; Whitfield, Z.J.; Dolan, P.T.; Sanchez-Vargas, I.; Garcia-Knight, M.; Ribiero, I.; Chen, T.; Olson, K.E.; Andino, R. Control of RNA viruses in mosquito cell through the acquisition of vDNA and endogenous viral elements. *Elife* **2019**, *8*, e41244. [CrossRef]

45. Liu, S.; Vijayendran, D.; Carrillo-Tripp, J.; Allen Miller, W.; Bonning, B.C. Analysis of new aphid lethal paralysis virus (ALPV) isolates suggests evolution of two ALPV species. *J. General Virol.* **2014**, *95*, 2809–2819. [CrossRef]