BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Current Trends

Volume 9(4)

Suppression subtractive hybridization (SSH) combined with bioinformatics method: an integrated functional annotation approach for analysis of differentially expressed immune-genes in insects

Chandan Badapanda

Interdisciplinary Research Center, Institute of Phytopathology & Applied Zoology, Justus-Liebig-University of Giessen, Heinrich-Buff-Ring 26-32, 35392 Giessen, Germany; Email: Chandan.Badapanda@agrar.uni-giessen.de

Received January 21, 2013; Accepted January 28, 2013; Published February 21, 2013

Abstract:
The suppression subtractive hybridization (SSH) approach, a PCR based approach which amplifies differentially expressed cDNAs (complementary DNAs), while simultaneously suppressing amplification of common cDNAs, was employed to identify immune-inducible genes in insects. This technique has been used as a suitable tool for experimental identification of novel genes in eukaryotes as well as prokaryotes; whose genomes have been sequenced, or the species whose genomes have yet to be sequenced. In this article, I have proposed a method for *in silico* functional characterization of immune-inducible genes from insects. Apart from immune-inducible genes from insects, this method can be applied for the analysis of genes from other species, starting from bacteria to plants and animals. This article is provided with a background of SSH-based method taking specific examples from innate immune-inducible genes in insects, and subsequently a bioinformatics pipeline is proposed for functional characterization of newly sequenced genes. The proposed workflow presented here, can also be applied for any newly sequenced species generated from Next Generation Sequencing (NGS) platforms.

Keywords: SSH, NGS, Immunity, Insects, Functional annotation, Bioinformatics.

Background:

Immune-related genes in insects

Insects have been remarkably successful in evolution. Current estimates are that they account for major species richness dominating any other class of biota **[1],** and relatively account for 80% of all animal species to date **[2].** To fight against infection or wounding, insect's immune system employ cellular and humoral immune system, although they do not possess adaptive immunity as present in case of higher vertebrates **[3-4].** Despite of this fact, insects' innate immune system is one of the ancient immune system constituting first line of defense, and conserved in different organisms, like in higher vertebrate, human.

Review on research progress using SSH and Bioinformatics

During these years, research on insects (physiology, biochemistry, and immunology) has produced voluminous information on the genome or transcriptome sequences of several insects, and bioinformatics has played a major role in the analysis, and management of the sequence data. There is remarkable progress in genome sequencing from insects after the first genome sequence of the fruit fly, *Drosophila*

melanogaster in the year 2000 **[5].** SSH-based sequencing technology has been the sequencing method used since 1960s **[6-7]** as a low cost alternatives for whole genome/ or transcriptome sequencing methods. The suppression subtractive hybridization (SSH) approach, a PCR based method which amplifies differentially expressed cDNAs, while simultaneously suppressing amplification of common cDNAs, was employed to identify immune-inducible genes in insects. This technique has been used as a suitable tool for experimental identification of novel genes in model insects whose genome has been sequenced such as the red flour beetle, *Tribolium castaneum* **[8],** or the pea aphid, *Acyrthosiphon pisum* **[9],** or the silkworm, *Bombyx mori* **[10].** In addition, and despite issues related to the sensitivity of this method, the SSH approach has been proven to be useful in targeted screening for immune-related genes in ancient insect, *Thermobia domestica*, for which no genomic data are available **[11],** or in insects which are adapted to habitats contaminated with high loads of microbes such as the rat-tailed maggots of the drone fly *Eristalis tenax* **[12]** capable of living in urine and cesspools, or the medicinal maggots of the green blow fly *Lucilia sericata* which reproduce in septic wounds or in carrions **[13]**, or the burying beetle *Nicrophorus vespilloides* **[4]** capable of living in cadaver rich with microbes and fungi. By the use of SSH-based approach, there are several other projects undertaken on different insects **[14-20]. Table 1 (see supplementary material)** represents a list of projects undertaken by different research groups in identifying immune-related genes in insects using SSH-based approach.

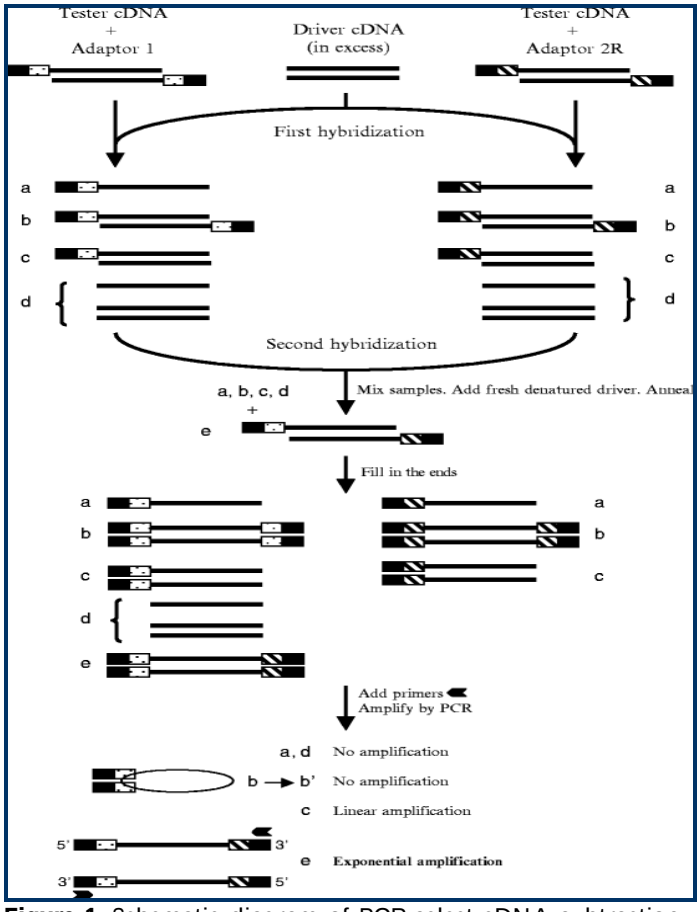

Figure 1: Schematic diagram of PCR-select cDNA subtraction. Tester and driver cDNAs were prepared from the total RNA of each treatment (control versus treatment). Suppression-subtractive hybridization was performed using PCR-Select cDNA Subtraction kit (Clontech Laboratories, Inc.). Forward and reverse subtraction libraries were constructed using cDNA samples of control versus treatment. The subtracted cDNAs were subjected to two rounds of PCR to normalize and enrich cDNA populations. The PCR products were sub-cloned into respective vector. **Type a** and **type d** molecules, devoid of primer annealing sites, cannot be amplified, **type b** molecules, being suppressed, cannot undergo the amplification process, **type c** molecules have only one primer site and can only be amplified linearly, and **type e** is differentially expressed, having only two adapters. This figure is adopted from **[7].**

Suppression subtractive hybridization (SSH)

In principle, the aim of the SSH-based sequencing approach is to build unbiased cDNAs library. The cDNA library will be the target of the subtraction (the tester population), which is denatured, and hybridized to another library that is present in the excess (the driver cDNAs population). Fragments common to both are annealed to each other, whereas tester specific products remain single stranded. Sequencing steps in PCR-select cDNA subtraction: The cDNA is synthesized from RNA generated from two types of cells being compared (tester and driver cDNAs); the tester and driver cDNAs are digested with RsaI to produce blunt ends; The tester DNA is divided into two parts, and each part is ligated with a different cDNA adaptor. As the ends of the adaptor do not contain a phosphate group, only one end of the adaptor attaches to the 5´ ends of the cDNA. Furthermore, the two adaptors contain a portion of identical sequence in order to allow annealing of the PCR primer once the recessed ends have been filled in. Two hybridization steps are performed. In the first step, an excess of driver is added to each sample of tester cDNA. "The samples are then heat denatured and allowed for annealing process and at last generating **a, b, c, d** types of molecules from each sample" **[7].** The concentration among high and low abundance sequences are equalized among **type a** molecules and at the same time **type a** molecules are significantly enriched with differentially expressed sequences, while cDNAs that are not differentially expressed in **type c** molecules hybridize with driver. In the second hybridization step, the two primarily hybridization samples are mixed together without denaturing. Through this, only the remaining equalized and subtracted single stranded tester cDNAs can hybridize, and only for **type e** molecule. This new hybrid is a double stranded tester molecule with different ends corresponding to **adaptor1** and **adaptor 2R**. Again, fresh denatured cDNA is added without denaturing the subtraction mix to enrich the fraction with differentially expressed cDNAs. After this event, the differentially expressed **type e** molecules are filled with DNA polymerase to make different annealing sites for the nested primers on their 5´ and 3´ ends. The whole population of molecules is subjected to PCR to amplify the differentially expressed sequences. During this process, **type a** and **type d** molecules cannot be amplified because they lack primer annealing sites. Due to the suppression effect, **type b** molecules cannot undergo the amplification process. **Type c** molecules have only one primer site and can only be amplified linearly. Only the **type e** molecules are equalized and differentially sequenced with two adaptors and can be amplified exponentially. (**Figure 1)** represents a schematic diagram of PCR-select cDNA subtraction based on SSH method. The overview of SSH method described above is

adopted from **[7],** therefore the readers are requested to read the above article for more details to set up the SSH-based experiment.

Bioinformatics methods for analysis of sequence data

Expressed Sequence Tags (ESTs) analysis

Expressed Sequence Tags (ESTs) are single-pass reads of approximately 200-800 base pairs (bp) generated from randomly selected cDNA clones. Since they represent the expressed portion of a genome, ESTs have proven to be extremely useful for gene identification and verification of gene predictions. They therefore represent a low-cost alternative to full genome sequences. (**Figure 2)** represents how the Expressed Sequence Tags (ESTs) are generated by SSH-based sequencing method.

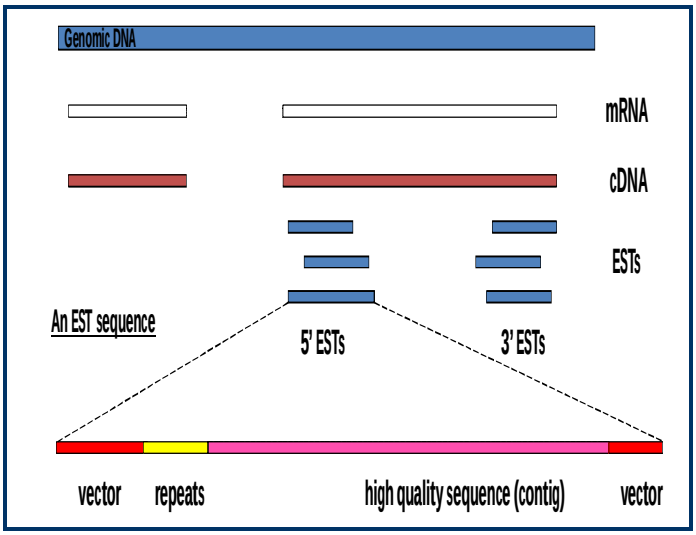

Figure 2: Graphical representation of Expressed Sequence Tags (ESTs). Expressed sequence tags, or ESTs, are single DNA sequencing reads made from complementary DNAs (cDNAs).Cellular mRNA is extracted from the tissue, it is reverse transcribed to produce DNA complementary to the initial mRNA, and incorporated into a plasmid, with appropriate vectors and linkers.

Transcriptome assembly

Bioinformatics tools are required to produce reliable, high quality data devoid of unwanted sequences in the pre-processing stage of EST projects. Pre-processing includes removing low-quality sequences, contaminant sequences (from vector to any other artifacts), and special features (like Poly-A tail or adaptors). Bioinformatics tools used in this process are LUCY2 **[21],** SeqClean (http://compbio.dfci.harvard.edu/tgi) and PreGap4 **[22].** Most of the transcriptomic studies for non-model organisms use the sequence assembly as a first step to generate contiguous sequences (contigs) that consist of overlapping reads to provide a consensus-based full length transcript. Multiple algorithms for *de novo* alignment have been developed, including: CAP3 **[23],** PHRAP (http://www.phrap.org/phredphrap/phrap.html), VELVET **[24]** and MIRA **[25].** Many assembler programs are available, which differ in details in their algorithmic development process. For example: CAP3 is based on Overlap-Layout Consensus (OLC) strategy, MIRA uses a hybrid-based method, whereas VELVET uses a graph-based approach. In terms of sensitivity, accuracy and memory requirement, CAP3 is memory intensive due to n^2 complexity of OLC-based method, whereas VELVET is less memory intensive as based on graph models **[26].** Assembly programs are also available commercially; a few of them are as follows: 'CLC Genomics Workbench', 'DNASTAR', 'AVADIS', etc. Transcriptome are assembled from shorter reads that vary in size, depending on the sequencing technology used. Contigs are created from these shorter reads by comparing all reads against each other. If the sequence identity and overlap length pass a certain threshold value, they are grouped together into a contig. For example, the following parameters have been used for processing transcriptomic datasets for the burying beetle, *Nicrophorus vespilloides* **[4].** Vector clipping, quality trimming and sequence assembly using stringent conditions (e.g. high quality sequence trimming parameters, 95% sequence identity cutoff, 25 bp overlap) was done with the Lasergene software package (DNASTAR Inc., Madison, WI; USA). (**Figure 3)** represents an overview of the EST analysis, which generates the consensus sequences.

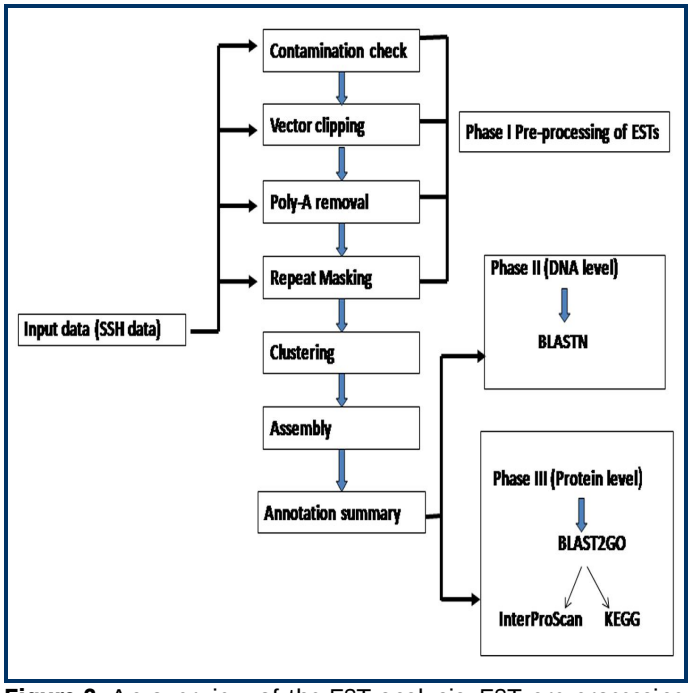

Figure 3: An overview of the EST analysis. EST pre-processing can be achieved by contamination check, vector sequence and poly-A removal, masking the repeats in the transcripts, followed by clustering and assembly of ESTs to generate contigs. Expressed Sequence Tags in FASTA format can be submitted to phase I for EST pre-processing. SeqClean (http://compbio.dfci.harvard.edu/tgi/) accepts ESTs in FASTA format, and performs vector removal, poly-A removal, trimming of low quality segments at the 5′ and 3′ ends and cleaning of low complexity regions. The output from SeqClean is processed by RepeatMasker (http://www.repeatmasker.org/) to mask repeats. CAP3 **[23]** then accepts repeat-masked high quality EST sequences and performs clustering and assembly into contigs. Alternatively, these pre-processing of ESTs can be achieved with DNASTAR (Medison USA) or CLC Genomics Workbench. The assembled ESTs may be submitted for functional annotation (phases II and III) either locally or through BLAST2GO suite tool.

Homology based detection

Homology based search is the best method for determining function by using NCBI Blastx algorithm, either installed locally to consume less time for large datasets, or achieved through the use of the NCBI online server **[27].** The Blastx program and associated databases may be downloaded for local Blast sequence searches (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/). Blast2GO offers a comprehensive suite tool for Blasting, with an advanced functional annotation **[28].** SwissProt protein database provides a high level of annotations, a minimal level of redundancy and a high level of integration with other databases, and thus is preferred for functional annotation. After a Blastx search, sequences may be compared to other nucleotide sequences with the program Blastn, or translated and compared to translate sequences using tBlastn to identify other unique sequences where Blastx did not work. However, Blastx is the first choice, since amino acid sequence is more conserved than nucleotide sequences. The statistical significant expectation value (E-value) predicts the probability that two sequences are related by chance. It is an important statistical parameter in performing Blast because setting E-value too low may create a false relationship, while setting high E-value presents the chance of excluding real sequences. As a consequence, by increasing the sequence length, the probability of finding significant Blast hits also increases. In real practice, performing Blast at low E-value and small sequence overlap length initially, and then filtering the results based on the distribution of hits obtained, may improve the obtainment of significant results from homology based approach. In the case of the *Nicrophorus* transcriptome analysis, first, sequences were searched against the NCBI non-redundant (nr) protein database using an E-value cut-off of 10^{-3}, with predicted polypeptides of a minimum length of 15 amino acids. Second, sequences retrieving no BLASTx hit were searched again by BLASTn, against an NCBI nr nucleotide database using an E-value cut-off of 10^{-10} **[4].**

Gene Ontology based functional annotation

Gene Ontology (GO) is a bioinformatics initiative that provides structural and controlled vocabulary of genes and genes products, to describe their cellular phenomena in terms of 'Biological Process', 'Molecular Function', and 'Cellular Components'. Along with this, the GO Consortium provides tools for easy access to all aspects of the data provided by the project **[29].** The interesting part of GO analysis is that GO vocabulary terms do not directly describe the gene or protein; rather, they describe the phenomenon of the gene or product of the gene. Gene Ontology annotation program is integrated with Blast2GO or can be used directly by Gene Ontology database, which can be accessed at http://www.geneontology.org/. Enzyme classification codes, and KEGG (Kyoto Encyclopedia of Genes and Genomes) metabolic pathway annotations, were generated from the direct mapping of GO terms onto their enzyme code equivalents by Blast2GO software program. (**Figure 3)** describes in details about the functional annotation of genes, which can be applied for functional characterization of genes from prokaryotes as well as eukaryotes.

Discussion:

Transcriptome of an organism is defined as its complete repertoire of transcripts, including its splice variants. SSH-based gene identification method was introduced as a cost-effective approach for rapid discovery and characterization of novel genes. mRNA expression varies- some are highly expressed in a cell while others are found in tiny amount. In whole transcriptomic sequencing project, a very large number of transcripts has to be sequenced, and to achieve this kind of goal, EST-project would be costly from an economic point of view. Thus for small scale EST-projects, SSH-based method is being used by different research laboratories around the globe. On the other hand, there are disadvantages in using SSH-based approach; like chances of missing rare transcripts, limited for gene expression studies, contigs generated are partial in length etc. In order to avoid missing rare transcript, normalization techniques have been devised to decrease the relative representation of the abundant transcript while increasing that of rare transcripts. SSH-based method is used to reduce the representation of transcripts already surveyed in previous libraries, as well as to enrich the sequences that are differentially expressed among specific tissues, cell types, etc. However, the possibility of missing rare transcripts by SSH-based method cannot be excluded, as in case of *Nicrophorus vespilloides*, we did not able to identify lysozyme **[4].** In recent times due to the popularity of high throughput sequencing techniques, SSH may be replaced by NGS technologies. The benefits of using RNA-Seq include high resolution, high dynamic range of expression (>8000 fold), low background noise, and the ability to identify allele specific expression and different isoforms **[30].** The use of sequence data from NGS technologies is considered as a powerful tool for differential gene expression analysis as well as for the fully comprehensive measurements of lower-abundance-class RNAs **[31-32].** Nevertheless, it is advisable to perform SSH-based approach for generating small scale cDNAs before investing into NGS technologies for whole transcriptome sequencing. To have a better understanding of the molecular and cellular mechanisms behind disease progression or biological development, the identification of differentially expressed genes has been important. So in recent times, the SSH-based method, have been used to identify the differential expressed genes linked to diseases or developmental process or to identify immune-inducible genes in insects upon challenged with microbes. Of note, the workflow described in this article is also applied for analysis of data derived from NGS technologies.

Acknowledgment:

I want to acknowledge my family, especially my father Mr. S. C. Badapanda, for his constant support and encouragement during my PhD Work. Also my special thanks to my friend Subhanjan Mohanty, who has supported me on this journey in so many different ways and on so many different levels.

References:

[1] Purvis A & Hector A, *Nature*. 2000 **405:** 212 [PMID: 10821281]
[2] Boman HG, *Annu Rev Immunol.* 1995 **13:** 61 [PMID: 7612236]
[3] Zou Z *et al. Genome Biol*. 2007 **8:** R177 [PMID: 17727709]
[4] Vogel H *et al. Insect Mol Biol*. 2011 **20:** 787 [PMID: 21929718]
[5] Celniker SE & Rubin GM, *Annu Rev Genomics Hum Genet*. 2003 **4:** 89 [PMID: 14527298]
[6] Bautz EK & Reilly E, *Science*. 1966 **151:** 328 [PMID: 5323418]
[7] Ghorbel MT & Murphy D, *Methods Mol Biol*. 2011 **789:** 237 [PMID: 21922412]

[8] Altincicek B *et al*. *Dev Comp Immunol*. 2008 **32:** 585 [PMID: 17981328]
[9] Altincicek B *et al. Insect Mol Biol*. 2008 **17:** 711 [PMID: 18823444]
[10] Bao YY *et al. Genomics*. 2009 **94:** 138 [PMID: 19389468]
[11] Altincicek B & Vilcinskas A, *Insect Biochem Mol Biol*. 2007 **37:** 726 [PMID: 17550828]
[12] Altincicek B & Vilcinskas A, *BMC Genomics*. 2007 **8:** 326 [PMID: 17875201]
[13] Altincicek B & Vilcinskas A, *Insect Mol Biol*. 2009 **18:** 119 [PMID: 19076250]
[14] Rinehart JP *et al. J Insect Physiol*. 2010 **56:** 603 [PMID: 20026067]
[15] Irles P *et al. BMC Genomics*. 2009 **10:** 206 [PMID: 19405973]
[16] Dixit R *et al. Int J Infect Dis*. 2009 **13:** 636 [PMID: 19128996]
[17] Zhao L *et al. J Med Entomol*. 2009 **46:** 490 [PMID: 19496418]
[18] Zhu Y *et al. Insect Biochem Mol Biol*. 2003 **33:** 541 [PMID: 12706633]
[19] Botha AM *et al. Plant Cell Rep*. 2006 **25:** 41 [PMID: 16328390]
[20] Wang Y *et al. Mol Plant Microbe Interact*. 2008 **21:** 122 [PMID: 18052889]
[21] Li S & Chou HH, *Bioinformatics*. 2004 **20:** 2865 [PMID: 15130926]
[22] Bonfield JK *et al. Nucleic Acid Res.* 1995 **23:** 4992 [PMID: 8559656]
[23] Huang X & Madan A, *Genome Res*. 1999 **9:** 868 [PMID: 10508846]
[24] Zerbino DR & Birney E, *Genome Res.* 2008 **18:** 821 [PMID: 18349386]
[25] Chevreux B *et al. Genome Res*. 2004 **14:** 1147 [PMID: 15140833]
[26] Kumar S & Blaxter ML, *BMC Genomics*. 2010 **11:** 571 [PMID: 20950480]
[27] Altschul SF *et al. J Mol Biol*. 1990 **215:** 403 [PMID: 2231712]
[28] Conesa A & Götz S, *Int J Plant Genomics*. 2008 **2008:** 619832 [PMID: 18483572]
[29] Gene Ontology Consortium, *Nucleic Acids Res*. 2010 **38:** D331 [PMID: 19920128]
[30] Deng X, *BMC Bioinformatics*. 2011 **12:** 267 [PMID: 21714929]
[31] Mortazavi A *et al. Nat Methods*. 2008 **5:** 621 [PMID: 18516045]
[32] Wilhelm BT & Landry JR, *Methods*. 2009 **48:** 249 [PMID: 19336255]

Edited by P Kangueane
Citation: Badapanda, Bioinformation 9(4): 216-221 (2013)

BIOINFORMATION

open access

Supplementary material:

Table 1: Projects undertaken using SSH-based method to identify novel genes in different species.

Species	Year	Genes	References
Nicrophorus vespilloides (burying beetle)	2011	1179	**[4]**
Tribolium castaneum (red flour beetle)	2008	75	**[8]**
Sarcophaga crassipalpis (flesh fly)	2010	97	**[14]**
Blatella germanica (German cockroach)	2009	258	**[15]**
Bombyx mori (silkworm)	2009	62	**[10]**
Anopheles stephensi (mosquito)	2009	32	**[16]**
Lucilia sericata (blow-fly)	2009	65	**[13]**
Acyrthosiphon pisum (pea aphid)	2008	35	**[9]**
Eristalis tenax (drone fly)	2007	30	**[12]**
Aedes aegypti (yellow fever mosquito)	2009	32	**[17]**
Thermobia domestica (Firebrat)	2007	26	**[11]**
Manduca sexta (tobacco hornworm)	2003	54	**[18]**
Diuraphis noxia (Russian wheat aphid)	2006	200	**[19]**
Nilaparvata lugens Stål (brown planthoppers)	2008	160	**[20]**