Discovery and analysis of RNA viruses in insects

Lumi Viljakainen and Jaana Jurvansuu

University of Oulu, Ecology and Genetics Research Unit, PO Box 3000, 90014 Oulu, Finland

Corresponding author e-mail: lumi.viljakainen@oulu.fi

Running Head: Discovery of RNA viruses

Summary

Viruses are the most abundant pathogens on Earth infecting all cellular life forms. Only in recent years have we started to gain knowledge on insect viromes, thanks to the development of sequencing technologies. The discovery and characterization of insect viruses is important for understanding insect-virus interactions, coevolution and insect immune defenses. We describe here a bioinformatic pipeline for the discovery of RNA viruses from insects based on RNA sequence data and the analysis of insect antiviral immune response against the discovered viruses by using small RNA sequence data.

Key words: Virus discovery, RNA viruses, insect viruses, bioinformatics pipeline, RNA-seq, sRNA-seq, small RNAs

1. Introduction

Insects hold the record for the most diverse animal taxon on the Earth (*1*) and in the recent years, with the aid of high-throughput sequencing methods, we have started to appreciate the rich RNA virome they harbor (*2-4*). Research on insect RNA viruses has an important value not only for increasing knowledge on virus biodiversity in the most numerous animal species but also to understand the role of insect microbiome in their biology. For example, in transcriptomic studies addressing questions related to insect immunity, it is noteworthy that the presence of RNA viruses may have an effect on inferences based on gene expression profiles (*5-7*).

We describe here the bioinformatics process of discovering replicating RNA viruses in insects based on two types of RNA sequence data: RNA-sequence (RNA-seq) and small RNA sequence (sRNA-seq) data. RNA-seq reads are typically 100-150 nucleotides (nt) in

length and optimal for the assembly of complete virus genomes. sRNA-seq reads are in maximum 50 nt in length and optimal for identifying viruses against which the host has raised an RNA interference response. When these short reads are mapped against the virus genomes, those viruses that are attacked by the host RNAi machinery will have a sRNA size distribution peaking at 21-22 nt indicative that the double-stranded RNA has been processed by the host Dicer enzyme.

It is noteworthy that the virome discovered with this process is dependent on the infection phase: non-replicating, latent viruses may remain undiscovered. Furthermore, different regions of the virus genome may be represented by varying amounts of sequencing reads due to differences in transcription levels of ORFs. This may lead to difficulties in the assembly of complete viral genomes. Any assembled virus genome represents a consensus sequence of many genotypes since RNA viruses have high mutation rates (8) and evolve as a quasispecies (9). Finally, sequences of DNA and retroviruses can be found from RNA-sequencing data (4). However, constructing a complete DNA virus sequence from RNA-sequencing data is difficult as DNA viruses are often large and not the whole genome is transcribed.

It is assumed that you already have RNA-seq and/or sRNA-seq data in fastq format on hand. You can find guidelines for experimental design to obtain such data in e.g. (10). The bioinformatic pipeline for the discovery and analysis of RNA viruses using RNA-seq data involves assembly of the sequencing reads to contigs, identification of virus-derived contigs using similarity search against a virus database, annotation of the virus genomes, quantification of the amount of reads derived from each virus, and phylogenetic analysis of the discovered viruses. In addition, if sRNA-seq data is available, it can be used for

determining against which of the viruses the host has an active immune response. You need to have the basic knowledge of Unix / Linux operating system to install the required software.

2. Materials

To be able to run the bioinformatics pipeline outlined below you need a computer with Unix/Linux operating system. The step that uses most computing power is the assembly of reads to contigs, which for example in Trinity, requires approximately 1GB of RAM per one million pairs of Illumina reads. The example command lines we present here do not take into account the possible sample-specific properties of your data, which might have to be addressed for example by additional or alternative parameters – you should always read the most recent documentation of each software you use.

3. Methods

3.1. Processing and analysis of RNA-seq data

1. The RNA-seq protocol here assumes you have strand-specific paired-end sequence data generated by the Illumina technology with read length of 100-150 nt. Make sure that your data consists of reads for which adapter sequences have been removed, low quality nucleotides have been trimmed from both ends of the reads and low quality reads have been discarded by using e.g. Trimmomatic (*II*) for cleaning the data:

```
trimmomatic PE -threads 8 -phred33 left_reads.fq.gz right_reads.fq.gz  
left_reads_paired.fq.gz left_reads_unpaired.fq.gz right_reads_paired.fq.gz  
right_reads_unpaired.fq.gz ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3  
TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

2. Check the quality and length distribution of the reads before and after trimming using FastQC (*12*).
3. Assemble the reads to contigs using Trinity (*13*):

```
Trinity --seqType fq --max_memory 380G --left left_reads_paired.fq --right  
right_reads_paired.fq --CPU 16 --output trinity_run_out
```

4. It is advisable to filter out contigs by a minimum length of e.g. 1000 nt, as the smallest RNA virus genomes are approximately 2000 nt (*14*). In this way, the manual work required for subsequent steps is reduced, while allowing some of the virus genomes to be partial:

```
bioawk -c fastx '{ if(length($seq) >= 1000) { print ">"$name; print $seq  
}}' Trinity.fasta > 1000_Trinity.fasta
```

5. Carry out a BLASTX similarity search with the contigs against the RefSeq release of viral protein sequences available at National Center for Biotechnology Information (NCBI (*15*)). It is advisable to download the viral sequences and use BLAST+, a stand-alone version of NCBI BLAST, on your local computer or a server. Download the virus protein sequences and name the file as virusproteins.fas and run BLASTX search. Here is an example command line:

```
blastx -query 1000_Trinity.fasta -db virusproteins.fas -outfmt "6 qseqid  
sseqid stitle qlen slen length evalue qcovs pident qstart qend" -evalue  
1e-10 -culling_limit 1 -out Blastx_1000_trinity_out.txt
```

An example of the BLASTX results is shown in Table 1.

6. In our experience, some of the Trinity-assembled virus contigs are derived from the same virus genome and may differ at only one or a few nucleotide positions – a very likely scenario given that viruses mutate at a high rate, and especially, if pooled insect samples have been used for preparing the RNA-seq library. Therefore, we recommend that the contigs that match viral proteins in the BLASTX search will be aligned using CAP3 (*16*) in order to merge contigs that differ due to small number (even only one) of mismatches, and additionally, to construct larger contigs based on overlaps that involve a few mismatches. Cap3 produces an output folder and five output files, from which .cap.contigs contains the new merged contigs and cap.singletons contains the contigs with no overlapping sequences with other contigs. Use both .cap.contigs and cap.singletons in further analysis.

cap3 virus_contigs_from_blastx.fas

7. Carry out a new BLASTX similarity search with the CAP3 contigs and singletons by following the instructions in step 5, and use the obtained information, i.e., subject titles of positive hits (stitle) together with their alignment lengths (length) and e-values (evaluate), for identification of virus-derived sequences and complete genomes. Furthermore, predict open reading frames (ORFs) with NCBI ORF Finder (*15*) or a similar tool, and perform another BLAST search to find a putative function for each ORF, which will help to confirm and annotate the virus

sequences. Note that many ORF still lack functional annotation, but protein level match in similarity search is an indication of conserved function.

8. Map the quality-filtered reads to the viral genomes to get insight on the abundance of each virus in your samples. When selecting the mapping software the read length must be taken into account. If the length of the majority of reads is >70 nt you can use e.g. BWA-MEM (*17*), if shorter, then e.g. Bowtie (*18*) is suitable. Here, we assume that the reads are > 70 nt and give example commands for BWA-MEM.

- a. First you need to index the virus genomes:

```
bwa index -a is_annotated_virus_genomes.fas
```

- b. Then, run the mapping:

```
bwa mem -t 2 annotated_virus_genomes.fas left_reads_paired.fq  
right_reads_paired.fq > output.sam
```

- c. Filter the output file according to mapping quality to discard reads that map to multiple locations. Especially viruses that have poly-A tails will have inflated read counts if the mapping quality filtering is skipped.

```
samtools view -q 20 -b output.sam > output_q20.bam
```

- d. To retrieve the read counts per virus genome, you first need to sort and index the virus genomes:

```
samtools sort output_q20.bam -o output_q20.sorted.bam
```

```
samtools index -b output_q20.sorted.bam
```

e. Finally, the read counts per each virus genome can be obtained:

```
samtools idxstats output_q20.sorted.bam > stats_per_virus_genome.tsv
```

9. To be able to compare viral read counts between samples you need to normalize the counts for the number of total filtered reads per sample and the length of the virus genome to get RPKM (reads per kilobase per million) counts. To obtain RPKM counts you need to divide the number of filtered reads by million to get a scaling factor. Then, divide the read count per virus genome first by the scaling factor (normalization for sequencing depth) and then by the length of the genome in kilobases.
10. It is advisable to set a threshold of viral read fraction out of the total number of filtered reads to be able to separate viruses that truly infect the insect host from those that might just be present in the digested food. For example, we have used a requirement of a minimum of 0.1% of reads mapping to the virus genome, after (3).
11. Active replication of positive-strand viruses in the host can be examined with strand-specificity of reads, as positive-strand viruses replicate through negative-strand intermediate. Strand-specificity information of reads derived from negative- and double-stranded viruses gives an indication of transcription efficacy of the genomes. Furthermore, as the double-stranded RNA is targeted by the RNA

interference response and, strand-specificity information can provide indication of how efficiently the host is able to react to the virus. The strand-specificity of the RNA-seq data can be analysed by running a perl script "examine_strand_specificity.pl" in the Trinity software package. This script examines the distribution of read orientations of the virus sequence. The script also draws an R-dependent violin plot.

```
TRINITY_HOME/util/misc/examine_strand_specificity.pl output_q20.sorted.bam
```

An example of the script output is shown in Table 2.

3.2. Processing and analysis of small RNA (sRNA) sequence data

1. The sRNA protocol assumes you have single-end size-selected (read length ≤ 50 nt) sequence data generated by Illumina technology. Again, remove adapter sequences, trim low quality nucleotides from both ends of the reads and discard low quality reads by using e.g. Trimmomatic (*11*):

```
trimmomatic SE -threads 8 -phred33 SE_reads.fq.gz filtered_SE_reads.fq.gz  
ILLUMINACLIP:TruSeq3-SE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15  
MINLEN:15
```

2. It is possible to use VirusDetect pipeline to assemble contigs using the filtered sRNA reads, perform a BLAST search with the contigs against specific virus databases and map the filtered reads against the virus contigs to get sRNA size distribution for each virus to identify viruses that have been targets of the host RNAi response (*19*). However, when we used both RNA-seq and sRNA-seq for

the same samples, we were able to assemble complete virus genomes only with the RNA-seq approach since the contigs assembled from sRNA reads were very short (mean length 180 nt). Therefore, we recommend that you use the VirusDetect pipeline if you have only sRNA-seq data. In case you have both RNA-seq and sRNA-seq data, assemble complete virus genomes with the RNA-seq data as above and continue with the next step.

3. Map the sRNA-seq reads against the virus genomes in order to get the sRNA size distribution for each virus. Here, the read length is the most important factor when selecting the mapping software. Now that the read length is < 50 nt, e.g. Bowtie (18) is a suitable software:

- a. First, index the virus genomes (the argument for the <ebwt_base> parameter is a user selected basename for the index files).

```
bowtie-build -q -f annotated_virus_genomes.fas <ebwt_base>
```

- b. Then, map the reads against the virus genomes:

```
bowtie --quiet -v 1 -a --best --strata -S <ebwt_base>  
filtered_SE_reads.fq.gz sRNA_output.sam
```

- c. Filter the output file for mapping quality:

```
samtools view -q 20 -b sRNA_output.sam -o sRNA_output_q20.bam
```

- d. Sort and index the quality-filtered file:

```
samtools sort sRNA_output_q20.bam -o sRNA_output_q20_sorted.bam
```

```
samtools index -b sRNA_output_q20_sorted.bam
```

4. To find out whether a virus has activated the host RNAi response, you need to visualize the size distribution of sRNAs that map to the virus. An R package ViRome (20) can be used for generating the size distribution and many other charts based on the bam file. In Figure 1a is a sRNA size distribution of an virus that shows a clear peak at 21 nt, both in the sense and antisense strands, indicating an active RNA interference response against the virus. Figure 1b is an example of sRNA size distribution without a peak at 21-22 nt and uneven representation of sense- and antisense-derived siRNAs suggesting that the virus is able to evade the host RNA interference response.

3.3. Phylogenetic analysis of the discovered viruses

Phylogenetic analysis, together with the information on virus genome organization, can be used to classify viruses and to investigate the relationship of a novel virus with other, similar viruses. The protein sequence of the RNA dependent RNA polymerase (RdRP) is typically used for the phylogenetic analysis since that is the most conserved protein of RNA viruses, although it should be noted that it gives a restricted picture of the evolutionary history as there is considerable amount of recombination among different viruses and even between viruses and cellular organisms (3). You may select the viruses for phylogenetic analysis by performing a BLASTP search with the RdRP protein sequence of the novel virus,

downloading the protein sequences of selected BLASTP search hits in fasta format and performing the following steps:

1. Multiple sequence alignment of the novel virus RdRP sequence with the selected BLASTX hits using MAFFT (21):

```
mafft --ep 0 --genafpair --maxiterate 1000 input.fas > alignment.fas
```

2. The alignments tend to contain regions that are difficult to align reliably, so it is recommended that these regions are trimmed with TrimAl (22):

```
trimal -in alignment.fas -out trimmed_alignment.fas -gt 0.5 -st 0.001 -  
cons 60 -sgc -scc -sident
```

3. Select amino acid substitution model using ProtTest (23):

```
java -jar prottest-3.4.2.jar -i trimmed_alignment.fas -o  
prottest_output.txt -log enabled -all-distributions -AICC -F
```

4. Reconstruct the phylogeny with PhyML (24) with the command below, where the arguments for the parameters in brackets are obtained in the ProtTest output:

```
phym1 -i trimmed_alignment.phy -d aa -m [model] -f [eq_freq] -v  
[invariable_sites] -a [gamma]
```

4. References

1. Berenbaum M (2017) Insect Biodiversity – Millions and Millions. In: Footitt RG, Adler PH (eds) Insect Biodiversity: Science and Society, Volume 1, 2nd edn. Wiley-Blackwell
2. Webster CL, Waldron FM, Robertson S et al (2015) The discovery, distribution, and evolution of viruses associated with *Drosophila melanogaster*. PLoS Biol 13(7):e1002210. <https://doi.org/10.1371/journal.pbio.1002210>
3. Shi M, Lin X, Tian J et al (2016) Redefining the invertebrate RNA virosphere. Nature 540:539-543
4. Viljakainen L, Holmberg I, Abril S et al (2018) Viruses of invasive Argentine ants from the European Main supercolony: characterization, interactions and evolution. J Gen Virol 99(8):1129-1140
5. Doublet V, Poeschl Y, Gogol-Döring A et al (2017) Unity in defence: honeybee workers exhibit conserved molecular responses to diverse pathogens. BMC Genomics 18:207 <https://doi.org/10.1186/s12864-017-3597-6>
6. Gerth M, Hurst GD (2017) Short reads from honey bee (*Apis* sp.) sequencing projects reflect microbial associate diversity. PeerJ 5:e3529. <https://doi.org/10.7717/peerj.3529>
7. Viljakainen L, Jurvansuu J, Holmberg I et al (2018) Social environment affects the transcriptomic response to bacteria in ant queens. Ecol Evol 8(22):11031-11070
8. Sanjuan R, Nebot MR, Chirico N et al (2010) Viral mutation rates. J Virol 84:9733-9748
9. Nowak MA (1992) What is a quasispecies? Trends Ecol Evol 7:118-121
10. Conesa A, Madrigal P, Tarazona S et al (2016) A survey of best practices for RNA-seq data analysis. Genome Biol 17:13. <https://doi.org/10.1186/s13059-016-0881-8>
11. Bolger AM, Lohse M Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120

12. Andrews S (2018) FASTQC, A quality control tool for high throughput sequence data.
<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. Accessed 18 April 2018.
13. Grabherr MG, Haas BJ, Yassour M et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29:644-652
14. Campillo-Balderas JA, Lazcano A Becerra A (2015) Viral genome size distribution does not correlate with the antiquity of the host lineages. *Front Ecol Evol* 3:143.
<https://doi.org/10.3389/fevo.2015.00143>
15. NCBI Resource Coordinators (2017) Database resources of the national center for biotechnology information. *Nucleic Acids Res* 45:D12-D17
16. Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. *Genome Res* 9:868-877
17. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754-1760
18. Langmead B, Trapnell C, Pop M et al (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25.
<https://doi.org/10.1186/gb-2009-10-3-r25>
19. Zheng Y, Gao S, Padmanabhan C et al (2017) VirusDetect: An automated pipeline for efficient virus discovery using deep sequencing of small RNAs. *Virology* 500:130-138
20. Watson M, Schnettler E Kohl A (2013) viRome: an R package for the visualization and analysis of viral small RNA sequence datasets. *Bioinformatics* 29:1902-1903
21. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30:772-780
22. Capella-Gutiérrez S, Silla-Martínez JM Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972-1973

23. Darriba D, Taboada GL, Doallo R et al (2011) ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27:1164-1165
24. Guindon S, Dufayard J, Lefort V et al (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59:307-321

Figure captions

Figure 1. The sRNA size distribution for two viruses, where the size is on the x-axis and the frequency on the y-axis. Sense-strand derived sRNAs are above the x-axis and antisense-strand derived sRNAs below the x-axis. 1a) An example of a distribution from a virus that has been recognized and processed by the host Dicer enzyme to generate mainly 22 nt long small interfering RNAs (siRNAs). The host Dicer recognizes double-stranded RNA, and therefore we expect to see equal representation of sense- and antisense-derived siRNAs. 1b) An example of a virus-derived sRNAs that most probably have been degraded and not processed by the host Dicer enzyme.

Table captions

Table 1. Example of BLASTX search results for two Trinity-contigs. The first contig matches several regions of *Lasius niger* virus 1, and the second contig matches a small region of Hubei picorna-like virus 59.

Table 2. Example of the strand-specific analysis results for viruses with negative- (-), positive- (+), or double-stranded (ds) RNA genomes.

Table 1. Example of BLASTX search results for two Trinity-contigs. The first contig matches several regions of *Lasius niger* virus 1, and the second contig matches a small region of Hubei picorna-like virus 59.

qseqid	sseqid	stitle	qle n	sl en	len gt h	eval ue	qc ov s	pid ent	qst art	qe nd
TRINITY_DN3	ref YP_00	RNA-dependent RNA	12							
3406_c0_g1_i	9407943.1	polymerase, partial [<i>Lasius</i>	04	21	21			86.	70	61
1		<i>niger</i> virus 1]	0	45	50	0	53	09	56	6
TRINITY_DN3	ref YP_00		12							
3406_c0_g1_i	9407942.1	putative capsid protein [<i>Lasius</i>	04	48	48			81.	91	77
1		<i>niger</i> virus 1]	0	7	8	0	12	56	53	17
TRINITY_DN3	ref YP_00		12			9.65				
3406_c0_g1_i	9407941.1	putative capsid protein [<i>Lasius</i>	04	27	27	E-		91.	99	91
1		<i>niger</i> virus 1]	0	0	0	171	7	11	65	56
TRINITY_DN3	ref YP_00		12			5.71			11	10
3406_c0_g1_i	9407938.1	putative capsid protein [<i>Lasius</i>	04	25	25	E-		92.	52	76
1		<i>niger</i> virus 1]	0	4	4	158	6	13	8	7
TRINITY_DN3	ref YP_00		12			7.68			10	
3406_c0_g1_i	9407939.1	hypothetical protein [<i>Lasius</i>	04	26	27	E-		85.	76	99
1		<i>niger</i> virus 1]	0	9	0	110	7	19	7	58
TRINITY_DN3	ref YP_00		12						10	10
3406_c0_g1_i	9407940.1	putative transmembrane	04			6.15		88.	76	50
1		protein [<i>Lasius niger</i> virus 1]	0	85	86	E-40	2	37	3	6
TRINITY_DN3	ref YP_00		10					25.	10	
3382_c1_g1_i	9337100.1	hypothetical protein [Hubei	85	29	13	4.22		95	57	67
1		picorna-like virus 59]	7	04	41	E-72	35	1	6	37

Table 2. Example of the strandedness analysis results with Trinity code for viruses with negative- (-), positive- (+), or double-stranded (ds) RNA genome.

#transcript	plus_strand_1stReads	minus_strand_1stReads	total_reads	diff_ratio
Virus1 (+)	61149	7220	68369	0,789
Virus2 (+)	53049	106	53155	0,996
Virus3 (-)	6448	1238	7686	0,678
Virus4 (-)	1080	2743	3823	-0,435
Virus5 (ds)	1302	398	1700	0,532