Virus detection with small RNA sequencing
Outline

- Small interfering RNA (siRNA) as an antiviral defense mechanism
- VirusDetect pipeline
  - central concepts
  - analysis steps
  - result files
RNA interference (RNAi)

- RNAi is an antiviral defense mechanism in eukaryotic organisms
  - Upon viral infection, Dicer enzymes cut viral dsRNA to small interfering RNA (siRNA) molecules, which are 21-24 nucleotides long.
  - siRNAs are further amplified by RNA-dependent RNA polymerases (RdRP).
  - siRNAs associate with Argonaute proteins and guide the RNA-induced silencing complex (RISC) to degrade viral RNA.

- We can sequence siRNAs and assemble virus genomes from the reads → virus detection and identification.
the cytoplasm

DNA virus (e.g. geminivirus)

nucleus

overlapping ssRNAs

replicating RNA virus

dsRNA (RNA virus)

"Degradative PCR"

RdRP

Systemic signal

Maintenance, amplification

Targeted RNA degradation

RISC nuclease complex

Complementary ssRNA or dsRNA

Targeted RNA degradation

Figure by Dr Jan Kreuze
Procedure

Extract RNA & run in 4% agarose gel

Cut and purify 20-30 nt band, send to sequencing provider for processing & sequencing on Illumina HiSeq 2000
VirusDetect – bioinformatics pipeline for detecting viruses in small RNA-seq data

- Combines several analysis steps in one tool
- Assembles RNA reads to longer sequences (contigs) in two ways
  - Reference-guided assembly: match reads to known virus sequences and combine matching reads together
  - de novo assembly: match reads to each other
- Compares the contigs to known virus sequences
  - Uses BLAST for similarity search
  - If no similarity is found, reports siRNA size distribution profile

- Read more
  - [http://virusdetect.feilab.net/cgi-bin/virusdetect/](http://virusdetect.feilab.net/cgi-bin/virusdetect/)
VirusDetect steps for contig assembly

1. sRNA reads
2. Alignment to host genome (host organism)
3. Unmapped sRNA reads
4. De novo assembly
5. Contigs
6. Alignment to host genome (host organism)
7. De novo contigs
8. Contigs (virusdetect_contigs.fa)
9. Virus reference based contigs
10. Alignment to virus reference database and reference-guided assembly
11. Host derived sRNA reads
12. Host derived contigs
VirusDetect steps for virus identification

- Contigs without hits
  - BLASTN against virus reference database. Result filtering*
    - BLASTN results (html, tsv, pdf, fasta, bam)
  - BLASTX against virus reference database. Result filtering*
    - BLASTX results (html, tsv, pdf, fasta, bam)

- Undetermined contigs
  - siRNA size profiling
    - Undetermined contigs (undetermined_contigs.fa, undetermined.html)

* Result filtering parameters:
  - Minimum fraction of a contig covered by virus reference (0.75)
  - Minimum fraction of virus reference covered by contigs (0.1)
  - Minimum read depth (5)
Reference-guided assembly

Virus reference sequence, 1000 nt

Assembled 2 contigs, coverage is 80% (800/1000 nt)
Reference-guided assembly of contigs

- Align reads with the aligner BWA to reference virus database
- As a reference database VirusDetect uses GenBank virus sequences. They have been
  - Classified to 8 different host kingdoms
    - Vertebrate, invertebrate, plant, protozoa, algae, fungus, bacteria, archaea
  - Processed to remove redundancy (sequences that are more than 95% similar have been combined)
  - The same reference database is used later on for BLAST searches

- Perform reference-guided assembly of the aligned reads using Samtools
De novo assembly of contigs

- Remove host-derived small RNAs
  - Align reads to host genome with BWA, keep the unaligned reads
- Assemble de novo contigs with Velvet
- Remove host-derived contigs
  - Align contigs to host genome with BWA
De novo assembly with Velvet

- Short words called k-mers are used to represent a sequence
  - k is the length of the word (5 in the example below)
- Velvet looks for overlaps between the words
- VirusDetect optimizes Velvet parameters
  - Runs Velvet using different k-mer lengths, selects the one that gives the best assembly (longest contigs), and then tries different coverage cutoffs with that k-mer length

Figure by Dr Jan Kreuze
Should host genome subtraction be used?

- Yes, because it improves the Velvet assembly and increases the efficiency of virus detection
  - enriches virus-derived siRNAs, reduces noise from host-derived small RNAs

- Yes, because some host genomes contain integrated viral sequences
  - related to extant replicating viruses but are mostly inactive fragments, and could be falsely identified as an infecting virus

- No, because host genome can have inadvertent viral sequence contamination
Combine contigs from the two assemblies

- Contigs from the reference-guided assembly and de novo assembly are combined
- Redundant contigs are removed using the Megablast assembler
  - Assembles two sequences into a contig using alignment information generated by the Megablast program
- Reads are aligned to the resulting non-redundant contigs and base errors are corrected
VirusDetect steps for virus identification

1. Contigs without hits
2. BLASTN against virus reference database. Result filtering
3. BLASTX against virus reference database. Result filtering
4. Undetermined contigs
5. siRNA size profiling
6. Undetermined contigs (undetermined_contigs.fa, undetermined.html)

* Result filtering parameters:
  - Minimum fraction of a contig covered by virus reference (0.75)
  - Minimum fraction of virus reference covered by contigs (0.1)
  - Minimum read depth (5)
Parameters for filtering BLAST results

- **Minimum fraction of a contig covered by virus reference**
  - The BLAST match must cover at least this fraction of the contig in order to make it significant for virus assignment. By default only contigs that match to the reference viruses for more than 75% of their length are considered.

- **Minimum fraction of virus reference covered by contigs**
  - Virus assignment is reported only if at least this percentage of the virus reference is covered with significant matches to contigs. The default is 10%.

- **Minimum read depth**
  - The average number of times each nucleotide of the reference sequence is covered by reads
How to visualize read depth?

- Map reads to the virus reference sequences using the Bowtie aligner
- Select the resulting BAM file and the reference sequence, and visualize them in the Genome Browser
  - Your genome is at the end of the list
  - The different viruses appear in the Chromosome list
VirusDetect result tables and log file

- **blastn_matching_references.html**
  - Table listing reference viruses that have matching contigs identified by BLASTN.

- **Virus.bn.pdf**
  - Detailed BLASTN result for each virus reference match.

- **undetermined.html**
  - Table listing the length, siRNA size distribution and 21-22nt percentage of undetermined contigs. Potential virus contigs (21-22 nt > 50%) indicated in green.

- **undetermined_blast.html**
  - Table listing contigs that have hits in the virus reference database but not assigned to any reference viruses because they did not pass the 3 filtering criteria.

- **blastn_matches.tsv**
  - Table listing contigs listing that have hits in the virus reference database.

- **Vd.log**
  - Info on number of contigs assembled and reads that aligned
Vd.log

process sample inputseq (total read: 50000)
[11/01/17 07:16:10] Align reads to reference virus sequence database
  10987 reads aligned
  62 unique contigs were generated
  30315 reads aligned
  51 contigs were assembled
  No host-derived contig was removed
  39 unique contigs were generated
[11/01/17 07:16:46] Remove redundancies in virus contigs
  100 contigs were assembled
  No host-derived contig was removed
  70 unique contigs were generated
[11/01/17 07:16:54] Virus identification
  2 viruses were identified by nucleotide similarity (BLASTN)
  No virus was identified by translated protein similarity (BLASTX)
  Contigs having enrichment of 21-22nt sRNAs were identified as potential virus sequences.

Please check undetermined.html

*****************************************************************************
Following output files were collected:

- **Sample_7_virusdetect_contigs.fa** Sequences of non-redundant contigs derived through reference-guided and de novo assemblies.
- **Sample_7_contigs_with_blastn_matches.fa** Sequences of contigs that match to virus references by BLASTN.
- **Sample_7_undetermined_contigs.fa** Sequences of contigs that do not match to virus references.
- **Sample_7_blastn_matching_references.html** Table listing reference viruses that have corresponding virus contigs identified by BLASTN. In addition, a pdf formatted report file is returned for each match.
- **Sample_7_blastn_matches.tsv** Table of BLASTN matches to the reference virus database.
- **Sample_7_undetermined.html** Table listing the length, siRNA size distribution and 21-22nt percentage of undetermined contigs. Potential virus contigs are indicated in green.
- **Sample_7_undetermined_blast.html** Table listing contigs having hits in the virus reference database but not assigned to any reference viruses because they did not meet the coverage or depth criteria.

```
total 320K
-rw-r--r-- 1 chipster chipster 24K Nov  1 07:17 Sample_7_blastn_matches.tsv
-rw-r--r-- 1 chipster chipster 2.7K Nov  1 07:17 Sample_7_blastn_matching_references.html
-rw-r--r-- 1 chipster chipster 14K Nov  1 07:17 Sample_7_contigs_with_blastn_matches.fa
-rw-r--r-- 1 chipster chipster 104K Nov  1 07:17 Sample_7_KJ534601.bn.pdf
-rw-r--r-- 1 chipster chipster 144K Nov  1 07:17 Sample_7_M72416.bn.pdf
-rw-r--r-- 1 chipster chipster 1.2K Nov  1 07:17 Sample_7_undetermined_blast.html
-rw-r--r-- 1 chipster chipster 129 Nov  1 07:17 Sample_7_undetermined_contigs.fa
-rw-r--r-- 1 chipster chipster 987 Nov  1 07:17 Sample_7_undetermined.html
-rw-r--r-- 1 chipster chipster 15K Nov  1 07:16 Sample_7_virusdetect_contigs.fa
```

Results have been collected to a single tar formatted archive file.
You can use tool: Extract .tar or .tar.gz file in Utilities folder to extract result files from the tar archive.
coverage = bases (percentage) of the reference that is covered by contigs
depth = the average number of times each reference base is covered by reads
depth norm = normalized depth: the average number of times each reference base is covered by reads, per million of total reads
% identity = the average percentage of sequence identity to the reference of all contigs aligned to that reference
% iden max = maximum percentage of sequence identity of the contigs to the reference
• Good match has high coverage and high depth
• Blue = virus sequence
• Red = contigs (lighter color indicates lower identity)
How does BLAST work?

- Produces a set of “words” (short, fixed-length sequences based on the query sequence) and scans the databases sequences for word matches
- When a word match to a database sequence is found, it is used to initiate a gap-free extension.
- gap-free extensions that achieve a certain score are used to seed a gapped extension
- Gapped extensions that achieve a specified score are saved and used as seeds for a gapped extension that also calculates the insertions and deletions and may use more sensitive parameters.
### Blastn_matches.tsv

<table>
<thead>
<tr>
<th>Contig_ID</th>
<th>Contig_Seq</th>
<th>Contig_Len</th>
<th>Hit_ID</th>
<th>Hit_Len</th>
<th>Genus</th>
<th>Description</th>
<th>Contig_start</th>
<th>Contig_end</th>
<th>Hit_start</th>
<th>Hit_end</th>
<th>Hsp_identity</th>
<th>E_value</th>
<th>Hsp_strand</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONTG12</td>
<td>GAAACTAAA...</td>
<td>6445</td>
<td>KJ54601</td>
<td>6445</td>
<td>potovirus</td>
<td>Potato virus X isolate SA-CIP, complete genome.</td>
<td>1</td>
<td>6445</td>
<td>1</td>
<td>6445</td>
<td>632/6447(98%)</td>
<td>0.0</td>
<td>1</td>
</tr>
<tr>
<td>CONTG32</td>
<td>TTTGAAAAACT..</td>
<td>5308</td>
<td>M72416</td>
<td>7568</td>
<td>potovirus</td>
<td>Potato virus X complete genome.</td>
<td>5</td>
<td>5308</td>
<td>1</td>
<td>5307</td>
<td>5119/5309(96%)</td>
<td>0.0</td>
<td>1</td>
</tr>
<tr>
<td>CONTG56</td>
<td>CATTACCTCC..</td>
<td>1397</td>
<td>M72416</td>
<td>7568</td>
<td>potovirus</td>
<td>Potato virus X complete genome.</td>
<td>2</td>
<td>1396</td>
<td>5234</td>
<td>6628</td>
<td>1361/1395(97%)</td>
<td>0.0</td>
<td>1</td>
</tr>
<tr>
<td>CONTG56</td>
<td>CATTACCTCC..</td>
<td>1397</td>
<td>M72416</td>
<td>7568</td>
<td>potovirus</td>
<td>Potato virus X complete genome.</td>
<td>73</td>
<td>1203</td>
<td>6436</td>
<td>7568</td>
<td>1104/1134(97%)</td>
<td>0.0</td>
<td>1</td>
</tr>
<tr>
<td>CONTG68</td>
<td>GCCTTGTGAA..</td>
<td>209</td>
<td>M72416</td>
<td>7568</td>
<td>potovirus</td>
<td>Potato virus X complete genome.</td>
<td>1</td>
<td>201</td>
<td>6435</td>
<td>6435</td>
<td>197/201(98%)</td>
<td>1e-101</td>
<td>1</td>
</tr>
<tr>
<td>CONTG68</td>
<td>GCCTTGTGAA..</td>
<td>209</td>
<td>M72416</td>
<td>7568</td>
<td>potovirus</td>
<td>Potato virus X complete genome.</td>
<td>1</td>
<td>201</td>
<td>7368</td>
<td>7568</td>
<td>197/201(98%)</td>
<td>1e-101</td>
<td>1</td>
</tr>
<tr>
<td>CONTG20</td>
<td>CAAAAGAGAG..</td>
<td>163</td>
<td>M72416</td>
<td>7568</td>
<td>potovirus</td>
<td>Potato virus X complete genome.</td>
<td>1</td>
<td>163</td>
<td>4165</td>
<td>4527</td>
<td>157/163(96%)</td>
<td>3e-74</td>
<td>1</td>
</tr>
<tr>
<td>CONTG79</td>
<td>CAGTCCACCT..</td>
<td>149</td>
<td>KJ54601</td>
<td>6445</td>
<td>potovirus</td>
<td>Potato virus X isolate SA-CIP, complete genome.</td>
<td>1</td>
<td>114</td>
<td>5098</td>
<td>5211</td>
<td>113/114(99%)</td>
<td>4e-58</td>
<td>1</td>
</tr>
<tr>
<td>CONTG79</td>
<td>CAGTCCACCT..</td>
<td>149</td>
<td>AF172259</td>
<td>6435</td>
<td>potovirus</td>
<td>Potato virus X complete genome.</td>
<td>1</td>
<td>149</td>
<td>5098</td>
<td>5246</td>
<td>140/150(93%)</td>
<td>2e-56</td>
<td>1</td>
</tr>
<tr>
<td>CONTG14</td>
<td>CAAAAGACAC..</td>
<td>128</td>
<td>KJ54601</td>
<td>6445</td>
<td>potovirus</td>
<td>Potato virus X isolate SA-CIP, complete genome.</td>
<td>1</td>
<td>120</td>
<td>3495</td>
<td>3614</td>
<td>115/120(95%)</td>
<td>5e-51</td>
<td>1</td>
</tr>
</tbody>
</table>

- **Hsp** = high scoring pair (the matching area of two sequences)
- **E_value** = expectation value: how many matches would have occurred at a given score by chance
<table>
<thead>
<tr>
<th>Contig</th>
<th>ID</th>
<th>Length</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
<th>26</th>
<th>27</th>
<th>28</th>
<th>29</th>
<th>30</th>
<th>31</th>
<th>32</th>
<th>33</th>
<th>21-22 (%)</th>
<th>Depth</th>
<th>Depth (Norm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONTIG15</td>
<td>184</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>502</td>
<td>618</td>
<td>13</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>97.31</td>
<td>135.15</td>
<td>54.06</td>
</tr>
<tr>
<td>CONTIG62</td>
<td>171</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>702</td>
<td>194</td>
<td>15</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>96.97</td>
<td>115.01</td>
<td>46.00</td>
</tr>
<tr>
<td>CONTIG31</td>
<td>137</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1722</td>
<td>606</td>
<td>22</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>98.56</td>
<td>366.15</td>
<td>146.46</td>
</tr>
<tr>
<td>CONTIG25</td>
<td>129</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>874</td>
<td>330</td>
<td>18</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>97.65</td>
<td>203.30</td>
<td>81.32</td>
</tr>
<tr>
<td>CONTIG74</td>
<td>121</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1643</td>
<td>320</td>
<td>11</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>99.04</td>
<td>347.02</td>
<td>138.81</td>
</tr>
<tr>
<td>CONTIG66</td>
<td>113</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>447</td>
<td>175</td>
<td>14</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>97.04</td>
<td>120.88</td>
<td>48.35</td>
</tr>
<tr>
<td>CONTIG27</td>
<td>90</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1507</td>
<td>297</td>
<td>65</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>96.16</td>
<td>441.59</td>
<td>176.64</td>
</tr>
<tr>
<td>CONTIG59</td>
<td>87</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>85</td>
<td>32</td>
<td>8</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>81.25</td>
<td>35.97</td>
<td>14.39</td>
</tr>
<tr>
<td>CONTIG9</td>
<td>82</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>9</td>
<td>16</td>
<td>80</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14.29</td>
<td>31.89</td>
<td>12.76</td>
</tr>
<tr>
<td>CONTIG45</td>
<td>78</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>409</td>
<td>127</td>
<td>6</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>98.53</td>
<td>148.38</td>
<td>59.35</td>
</tr>
</tbody>
</table>

- Green color indicates potential virus contigs (21-22nt >50%)
VirusDetect sequence files and BAM

- **virusdetect_contigs.fa**
  - Sequences of non-redundant contigs derived through reference-guided and \textit{de novo} assemblies.

- **contigs_with_blastn_matches.fa**
  - Sequences of contigs that match to virus references by BLASTN.

- **undetermined_contigs.fa**
  - Sequences of contigs that do not match to virus references.

- **blastn_matching_references.fa and .fai.**
  - Virus reference sequences that produced BLASTN hits, and fasta index file

- **blastn_matches.bam and .bai.**
  - BAM file containing the BLASTN alignment of each contig to its corresponding virus reference sequence, and BAM index file.
How many reads should I have?

More reads give better reference coverage, longer contigs and more depth

Figure from Zheng Y et al (2017) VirusDetect: An automated pipeline for efficient virus discovery using deep sequencing of small RNAs. Virology 500:130-138
What kind of reads to use for VirusDetect?

- 50 bp, single end sequencing
- Check the quality with FastQC
- Trim away sequencing adapters
- Filter out reads that
  - contain Ns
  - are shorter than 15 nt