VirusDetect

Introduction

Accurate detection and identification of virus infection in plants and animals is critical for agriculture production and human health. Conventional methods such as PCR or microarrays are useful but they require the prior knowledge and sequence information of the potential pathogens, thus they are not efficient in detecting novel or emerging viruses. RNA silencing constitutes a fundamental antiviral defense mechanism in both plants and animals in which host enzymes cut and amplify viral RNA into pieces of 20-24 nt. Deep sequencing of these virus-derived small RNAs (sRNAs) and proper assembly or alignment of these sRNA sequences can reconstruct genomic sequence information of the viruses being targeted in the plants and animals. This approach is independent of the ability to culture or purify the virus and does not require any specific amplification or enrichment of viral nucleic acids as it automatically enriches for small RNAs of viral origin by tapping into a natural antiviral defense mechanism.

VirusDetect is a software package that can efficiently and exhaustively analyze large-scale sRNA datasets for virus identification. The program performs reference-guided assembly by aligning sRNA reads to the known virus reference database (GenBank gbvrl) as well as *de novo* assembly using Velvet with automated parameter optimization. The assembled contigs are compared to the reference virus sequences for virus identification.

System requirement and dependencies

- 64-bit Linux system - Mac OS X is not supported
- Perl version 5.10.0 or higher. Perl is installed by default on most Linux systems
- BioPerl version 1.006 or higher. Please check http://www.bioperl.org and wiki/Installing_BioPerl for more details on installation of BioPerl.
- BWA 0.7.10. Provided in VirusDetect.
- SAMtools v0.1.18. Provided in VirusDetect.
- Velvet v1.1.07. Provided in VirusDetect.
- NCBI BLAST package 2.2.16. Provided in VirusDetect.
- HISAT (for RNA-Seq datasets)

Download

Current version of VirusDetect is v1.5. It's available for 64-bit linux systems. Download VirusDetect from the ftp server

Installation

Installation of VirusDetect is straightforward. Download VirusDetect and unzip the downloaded file.

```
$ tar -xzvf VirusDetect_v1.5.tar.gz
```

This will generate a folder named "VirusDetect-v1.4" (we call this folder "VirusDetect home folder"). VirusDetect home folder includes three subfolders, a "bin" folder which contains all executables, a "databases" folder which holds the reference virus sequence and the host genome sequence databases, and a "tools" folder which provides the virus classification script and the sRNA processing script. The home folder also contains a perl script, VirusDetect.pl, which is the core script to run the VirusDetect pipeline.

Run VirusDetect

Quick Start

1. Put sRNA sequence files in fasta or fastq format into VirusDetect home folder
2. Go to VirusDetect home folder and run VirusDetect with the following command

   ```
   $ perl VirusDetect.pl input1 input2 ......
   ```

3. The program can take multiple files (input1, input2 .....) as the input and run the files one by one sequentially. The program will generate an output folder named such as **result_input1** for each input file which contains all the output files. See below for the description of the output files.

Input files

VirusDetect takes one or more sequence files in fasta or fastq format as its input. It's highly recommended to remove ribosomal RNA (rRNA) sequences from the input sequences before running VirusDetect. Users can align sRNA reads to the Silva rRNA database using bowtie. Here is the command we recommend (assuming the sRNA sequence file is in fasta format):

```
$ bowtie -v 1 -k 1 --un cleaned_sRNA -f -p 15 Silva_rRNA_database
sRNA_sequences sRNA_rRNA_match
```

Build virus reference and host genome databases

The virus reference database is available from GenBank (gbvrl). We classified these virus sequences into different kingdoms including plant, vertebrate, invertebrate, fungus, bacteria, algae, archaea and protozoa using the Virus Classification Pipeline we have developed. Unique virus sequence databases were generated for each host kingdom by removing redundant sequences of 100%, 97% and 95% identity, respectively. The classified and non-redundant databases are available on our ftp site. The classified virus sequence databases (both nucleotides and proteins) need to be properly built before used as the reference (the formatted databases are also provide on our ftp site):

```
$ bin/bwa index databases/known_virus_reference_nt
$ bin/formatdb -i databases/known_virus_reference_nt -p F
$ bin/formatdb -i databases/known_virus_reference_prot -p T
```

The host reference sequence databases, if available, can be used to subtract sRNA reads derived from the host. The databases also need to be properly built:

```
$ bin/bwa index databases/name_of_host_reference
```

Note: The virus reference and the host sequence databases must be put in the "databases" folder under the "VirusDetect home folder". A curated non-redundant plant virus sequence database (vrl_plant) is provided with the VirusDetect package. Virus sequence databases for other kingdoms can be obtained from our ftp site.

Parameters

```
$ perl virus_detect.pl --reference [FILE] [options] input1 input2 ......
```

Section 1: Basic parameters

--reference	**[String]**	Name of the reference virus database [vrl_plant]
--host_reference	**[String]**	Name of the host reference database used for host sequence subtraction [none]
--thread_num	**[Integer]**	Number of CPUs used for alignments [8]

Section 2: BWA alignment parameters (alignments of reads to reference viruses or host sequences)

--max_dist	**[Integer]**	Maximum edit distance [1]
--max_open	**[Integer]**	Maximum number of gap opens [1]
--max_extension	**[Integer]**	Maximum number of gap extensions [1]
--len_seed	**[Integer]**	Seed length [15]
--dist_seed	**[Integer]**	Maximum edit distance in the seed [1]

Section 3: HISAT options (align RNA-Seq reads to host references)

--hisat_dist	**[Integer]**	Maximum edit distance for HISAT [5]

Section 4: blast alignment options (remove redundancy within virus contigs)

--min_overlap	**[Integer]**	Minimum overlap length [30]
--max_end_clip	**[Integer]**	Maximum length of end clips [6]
--min_identify	**[Float]**	Minimum percent identity [97]
--mis_penalty	**[Integer]**	Penalty score for a nucleotide mismatch [-3]
--gap_cost	**[Integer]**	Cost to open a gap [-1]
--gap_extension	**[Integer]**	Cost to extend a gap [-1]

Section 5: blast alignment options (align virus contigs to virus reference database)

--word_size	**[Integer]**	Minimum word size [11]
--exp_value	**[Float]**	Maximum e-value [1e-5]
--identity_percen	**[Float]**	Minimum percentage identity [25]
--mis_penalty_b	**[Integer]**	Penalty score for a nucleotide mismatch [-3]
--gap_cost_b	**[Integer]**	Cost to open a gap [-1]
--gap_extension_b	**[Integer]**	Cost to extend a gap [-1]

Section 6: result filter options

--hsp_cover	**[Float]**	Coverage cutoff of a reported virus contig by reference virus sequences [0.75]
--coverage_cutoff	**[Float]**	Coverage cutoff of a reported virus reference sequences by assembled virus contigs [0.1]
--depth_cutoff	**[Float]**	Depth cutoff of a reported virus reference [5]

Output files

VirusDetect generates the following files in the output directory.

1. **contig_sequences.fa, contig_sequences.blastn.fa, contig_sequences.blastx.fa, and contig_sequences.undetermined.fa**

contig_sequences.fa:	Sequences of non-redundant contigs derived through reference-guided and *de novo* assemblies
contig_sequences.blastn:	Sequences of contigs that match to virus references by BLASTN
contig_sequences.blastx:	Sequences of contigs that match to virus references by BLASTX
contig_sequences.undetermined.fa:	Sequences of contigs that do not match to virus references.

2. **blastn.references.fa and blastx.references.fa**

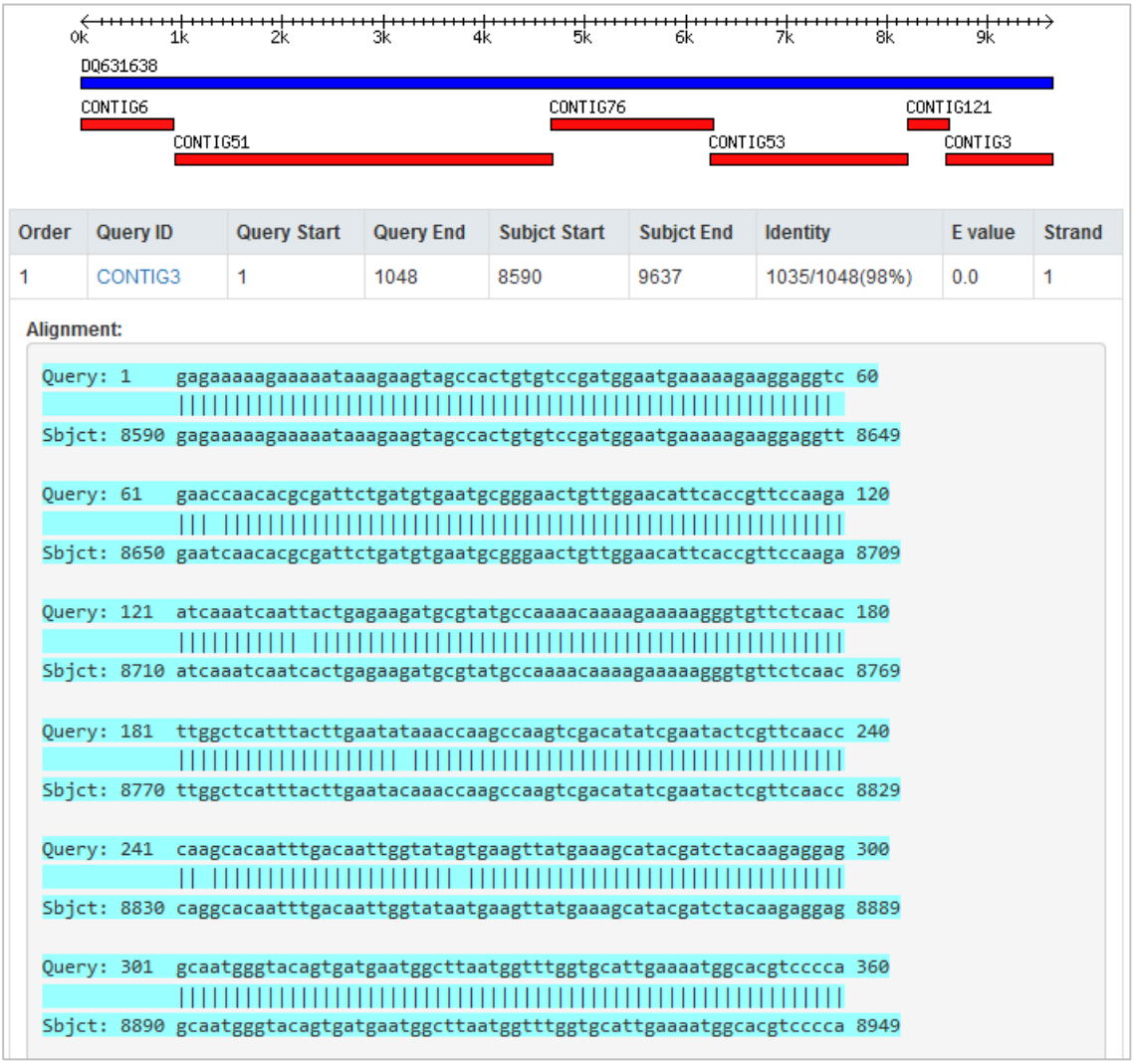
   The reference virus sequences that have corresponding aligned non-redundant contigs by BLASTN or BLASTX.

3. **blastn.html and blastx.html**

   The html files listing reference viruses that have corresponding virus contigs identified by BLASTN and BLASTX, respectively

Reference	Length	Coverage (%)	#contig	Depth	Depth (Norm)	%Identity	%Iden Max	%Iden Min	Genus	Description
KP223323	5858	5858 (100)	3	4526.5	1193.4	99.97	100	97.01	comovirus	Squash mosaic virus segment RNA-1, complete sequence.
KP223324	3370	3370 (100)	2	3915.0	1032.2	100	100	100	comovirus	Squash mosaic virus segment RNA-2, complete sequence.

The link of the accession number of each reference virus provides detailed alignments of virus contigs to the reference virus



Order	Query ID	Query Start	Query End	Subjct Start	Subjct End	Identity	E value	Strand
1	CONTIG3	1	1048	8590	9637	1035/1048(98%)	0.0	1

Alignment:

```
Query: 1    gagaaaaagaaaaataaagaagtagccactgtgtccgatggaatgaaaaagaaggaggtc 60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 8590 gagaaaaagaaaaataaagaagtagccactgtgtccgatggaatgaaaaagaaggaggtt 8649

Query: 61   gaaccaacacgcgattctgatgtgaatgcgggaactgttggaacattcaccgttccaaga 120
            ||| ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 8650 gaatcaacacgcgattctgatgtgaatgcgggaactgttggaacattcaccgttccaaga 8709

Query: 121  atcaaatcaattactgagaagatgcgtatgccaaaacaaaagaaaaagggtgttctcaac 180
            |||||||||||| ||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 8710 atcaaatcaatcactgagaagatgcgtatgccaaaacaaaagaaaaagggtgttctcaac 8769

Query: 181  ttggctcatttacttgaatataaaccaagccaagtcgacatatcgaatactcgttcaacc 240
            |||||||||||||||||||| ||||||||||||||||||||||||||||||||||||||
Sbjct: 8770 ttggctcatttacttgaatacaaaccaagccaagtcgacatatcgaatactcgttcaacc 8829

Query: 241  caagcacaatttgacaattggtatagtgaagttatgaaagcatacgatctacaagaggag 300
            || |||||||||||||||||||| ||||||||||||||||||||||||||||||||||||
Sbjct: 8830 caggcacaatttgacaattggtataatgaagttatgaaagcatacgatctacaagaggag 8889

Query: 301  gcaatgggtacagtgatgaatggcttaatggtttggtgcattgaaaatggcacgtcccca 360
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 8890 gcaatgggtacagtgatgaatggcttaatggtttggtgcattgaaaatggcacgtcccca 8949
```

4. **blastn.sam and blastx.sam**

SAM format files containing the alignment information of each contig to its corresponding virus reference sequences. The file can be viewed by Tablet, IGV, and many others...

5. **blastn.xls and blastx.xls**

Excel files containing the detailed alignment information between virus contigs to their corresponding virus reference sequences.

Contig_ID	Contig_Seq	Contig_Len	Hit_ID	Hit_Len	Genus	Description	Contig_start	Contig_end	Hit_start	Hit_end	Hsp_identity	E_value	Hsp_strand
CONTIG5	GAAAACAAA	6413	AF484251	6450	Potexvirus	Pepino mosaic virus isolate Sp-13, complete genome.	7	6413	2	6410	6343/6409(98%)	0	1
CONTIG10	AAGACGGCAT	90	AF484251	6450	Potexvirus	Pepino mosaic virus isolate Sp-13, complete genome.	13	90	2056	2133	78/78(100%)	1.00E-25	1
CONTIG33	AGGATCAACC	712	AF484251	6450	Potexvirus	Pepino mosaic virus isolate Sp-13, complete genome.	11	711	57	757	697/701(99%)	0	-1
CONTIG14	AGCTTGGTTA	330	AY509926	6413	Potexvirus	Pepino mosaic virus strain US1, complete genome.	1	330	6081	6410	320/330(96%)	1.00E-124	1
CONTIG28	GAAAACAAA	4827	AY509926	6413	Potexvirus	Pepino mosaic virus strain US1, complete genome.	1	4827	1	4828	4728/4829(97%)	0	1
CONTIG29	TCAACTTCAA	1244	AY509926	6413	Potexvirus	Pepino mosaic virus strain US1, complete genome.	1	1244	4834	6077	1229/1244(98%)	0	1
CONTIG3	CCCCTCGCCC	358	HM107843	360	unclassified viroids	Rubber viroid India/2009 isolate KER1, complete genome	48	355	1	308	303/310(97%)	1.00E-115	1
CONTIG11	CAGAACTAA	126	HM107843	360	unclassified viroids	Rubber viroid India/2009 isolate KER1, complete genome	3	124	90	209	117/122(95%)	3.00E-38	1
CONTIG24	ATTCGCTCCA	157	HM107843	360	unclassified viroids	Rubber viroid India/2009 isolate KER1, complete genome	26	156	176	307	130/132(98%)	8.00E-46	1

Contact

For questions and suggestions, please contact us at bioinfo@cornell.edu