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**

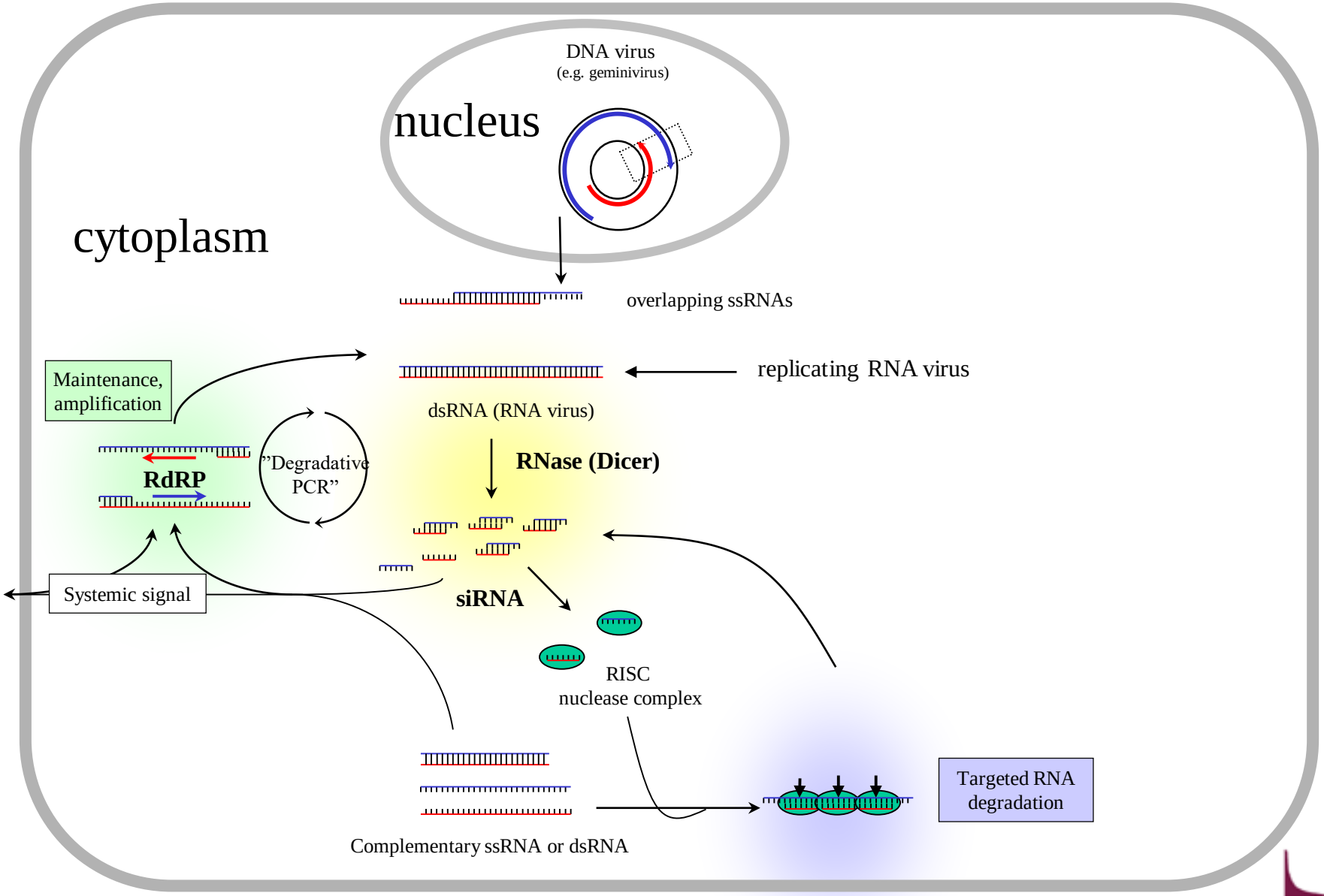


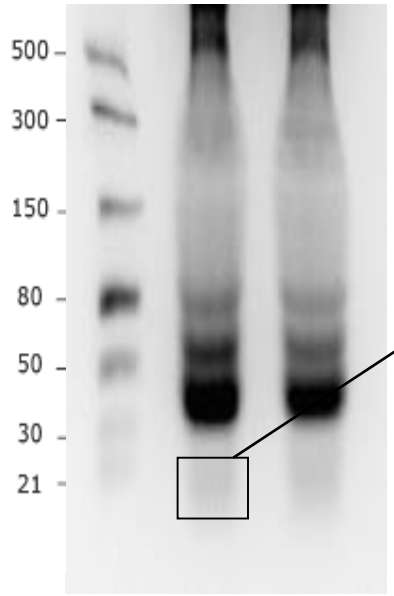
Figure by Dr Jan Kreuze

CSC

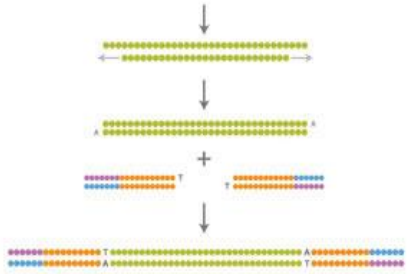
Procedure

Figure by Dr Jan Kreuze

Extract RNA & run in 4% agarose gel



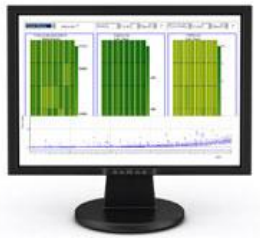
Cut and purify 20-30 nt band, send to sequencing provider for processing & sequencing on Illumina HiSeq 2000



Library Preparation

Cluster Generation

Sequencing by Synthesis



RTA v1.7, CASAVA v1.7

CSC

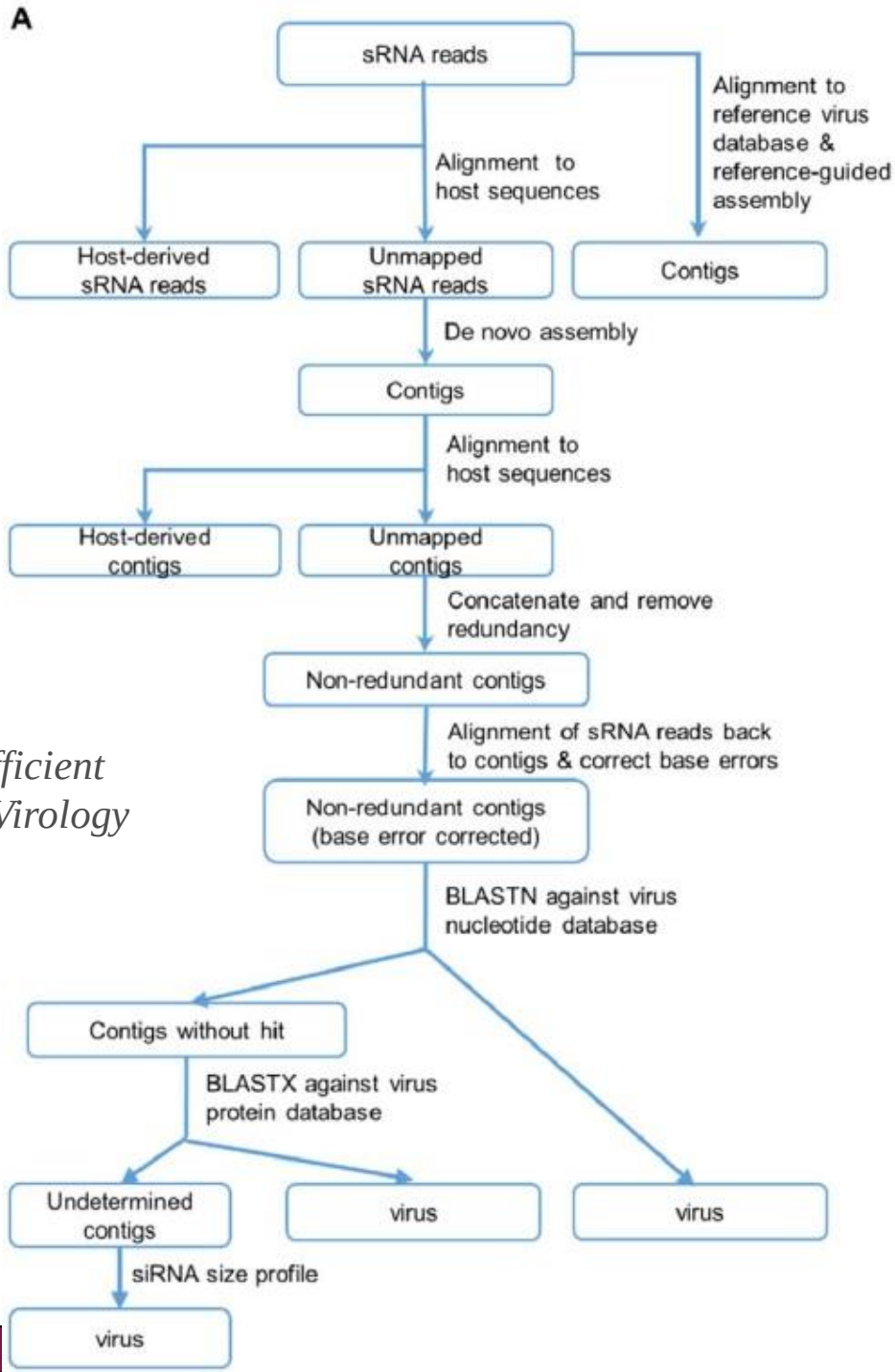
VirusDetect – bioinformatics pipeline for detecting viruses in small RNA-seq data

- **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**
  - Zheng Y et al (2017) VirusDetect: An automated pipeline for efficient virus discovery using deep sequencing of small RNAs. Virology 500:130-138
  - http://virusdetect.feilab.net/cgi-bin/virusdetect/

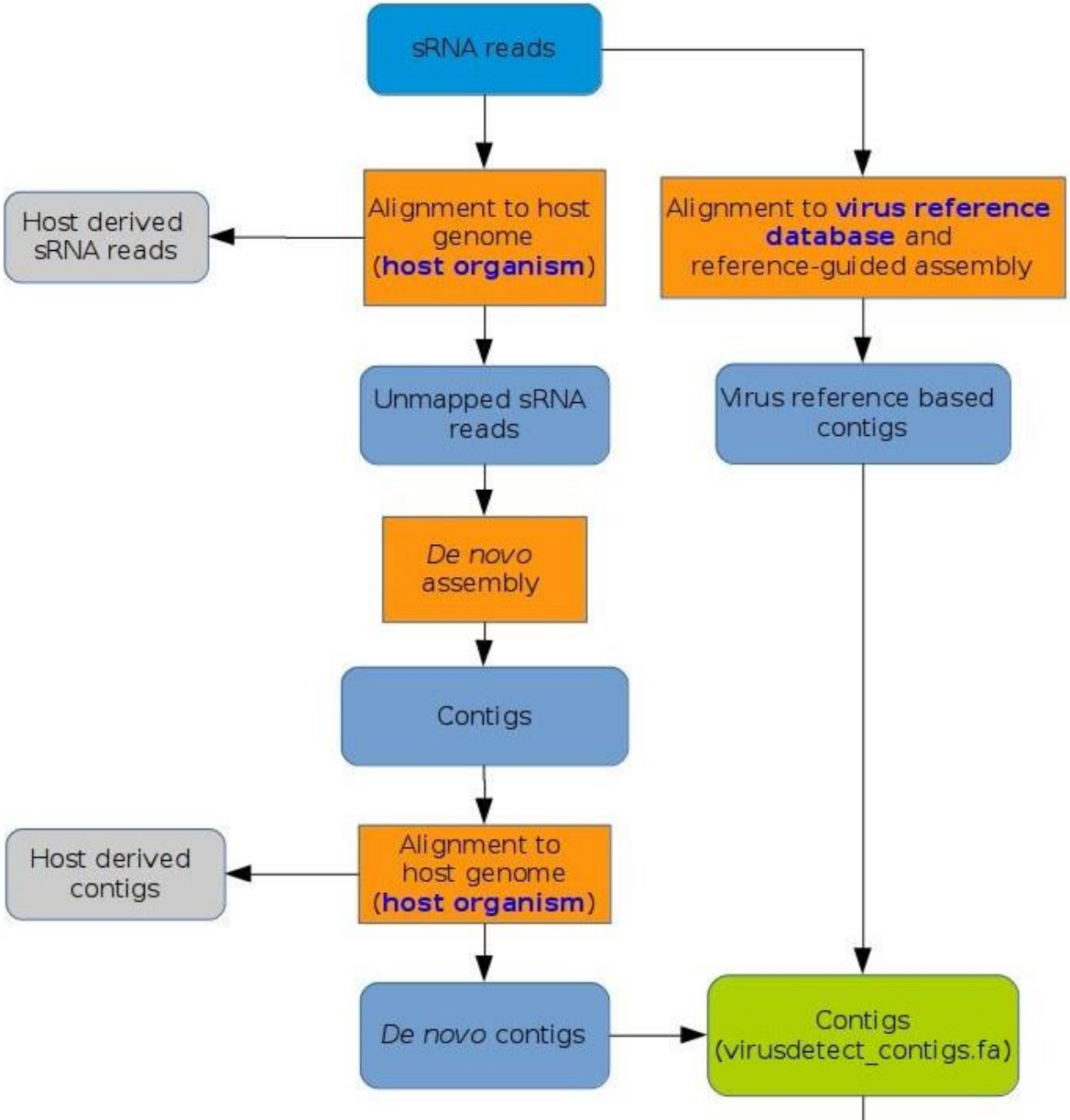
CSC

VirusDetect pipeline



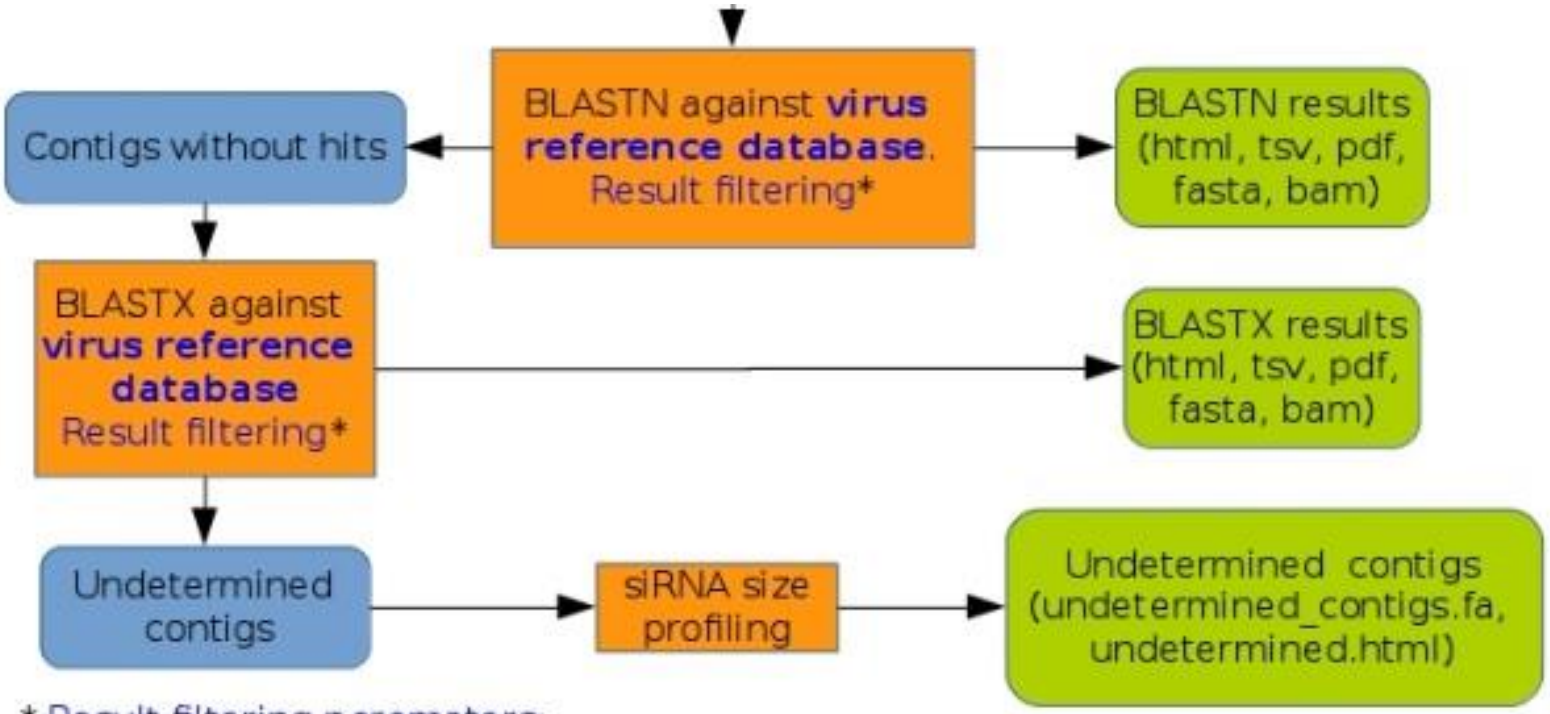
Zheng et al: VirusDetect – an automated pipeline for efficient virus discovery using deep sequencing of small RNAs. Virology 2017 (500) 131-138

VirusDetect steps for contig assembly



CSC

VirusDetect steps for virus identification

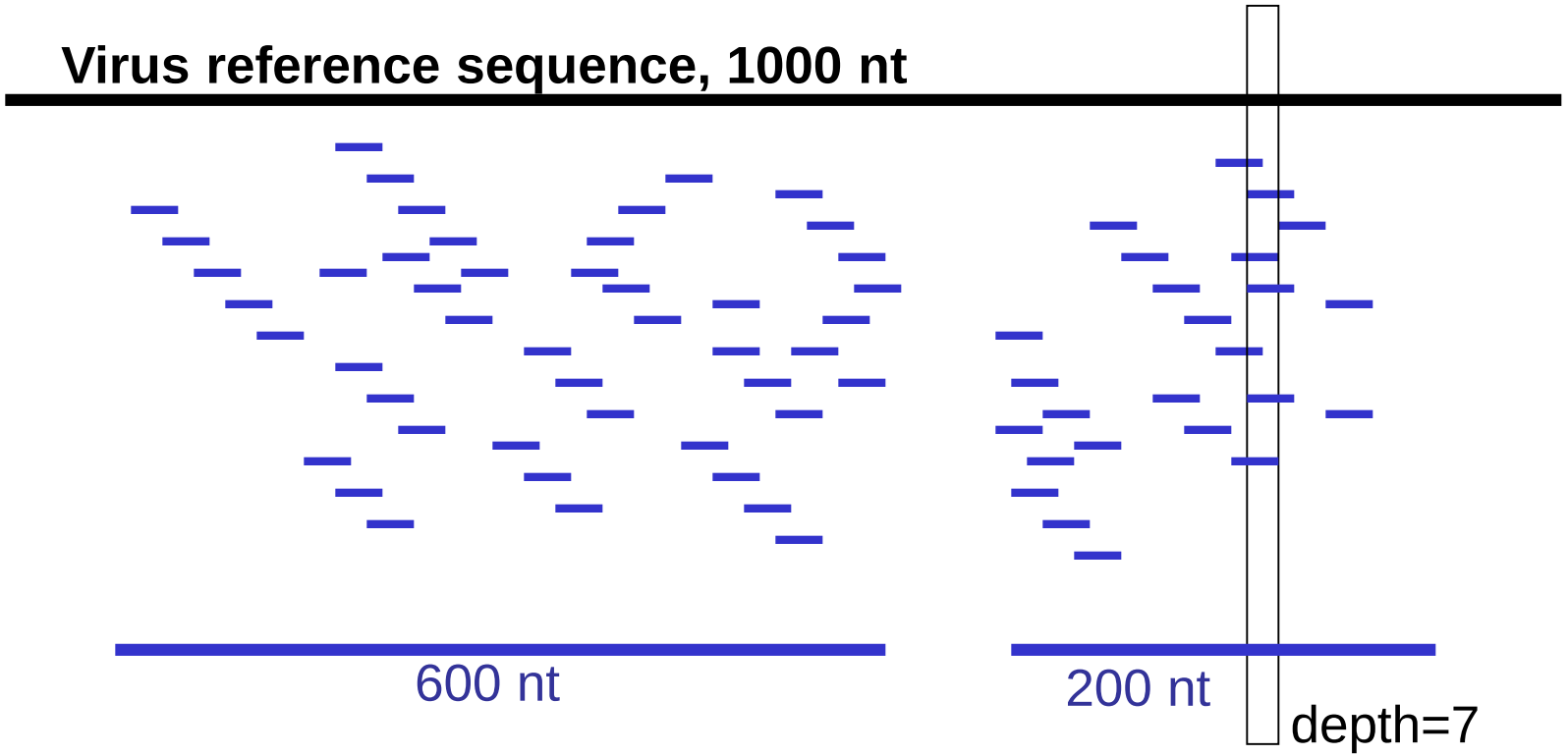


* Result filtering parameters:
- Minimum fraction of a contig covered by virus reference (0.75)
- Minimum fraction of virus reference covered by contigs (0.1)
- Minimum read depth (5)

CSC

Reference-guided assembly

Figure by Dr Jan Kreuze



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

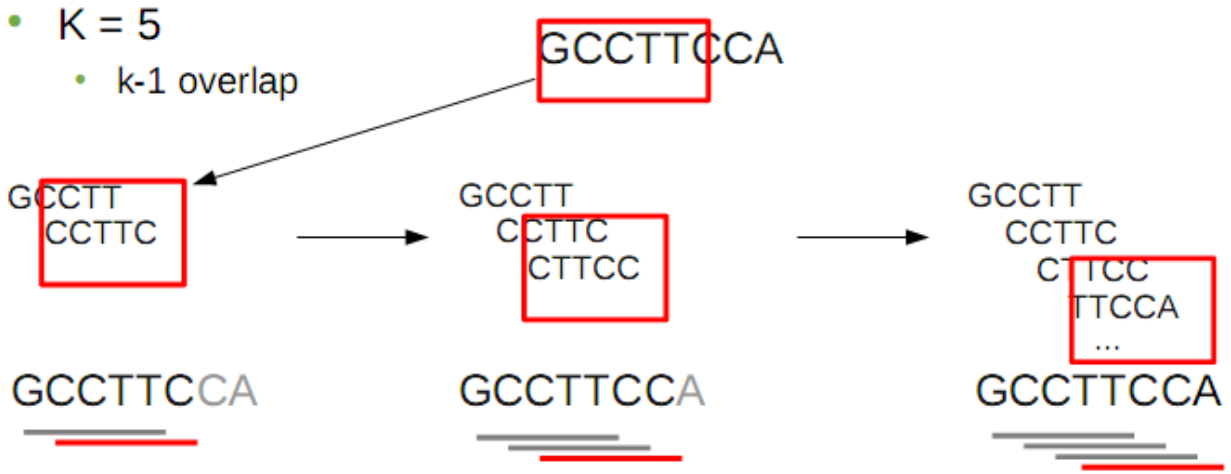


Figure by Dr Jan Kreuze

Should host genome subtraction be used?

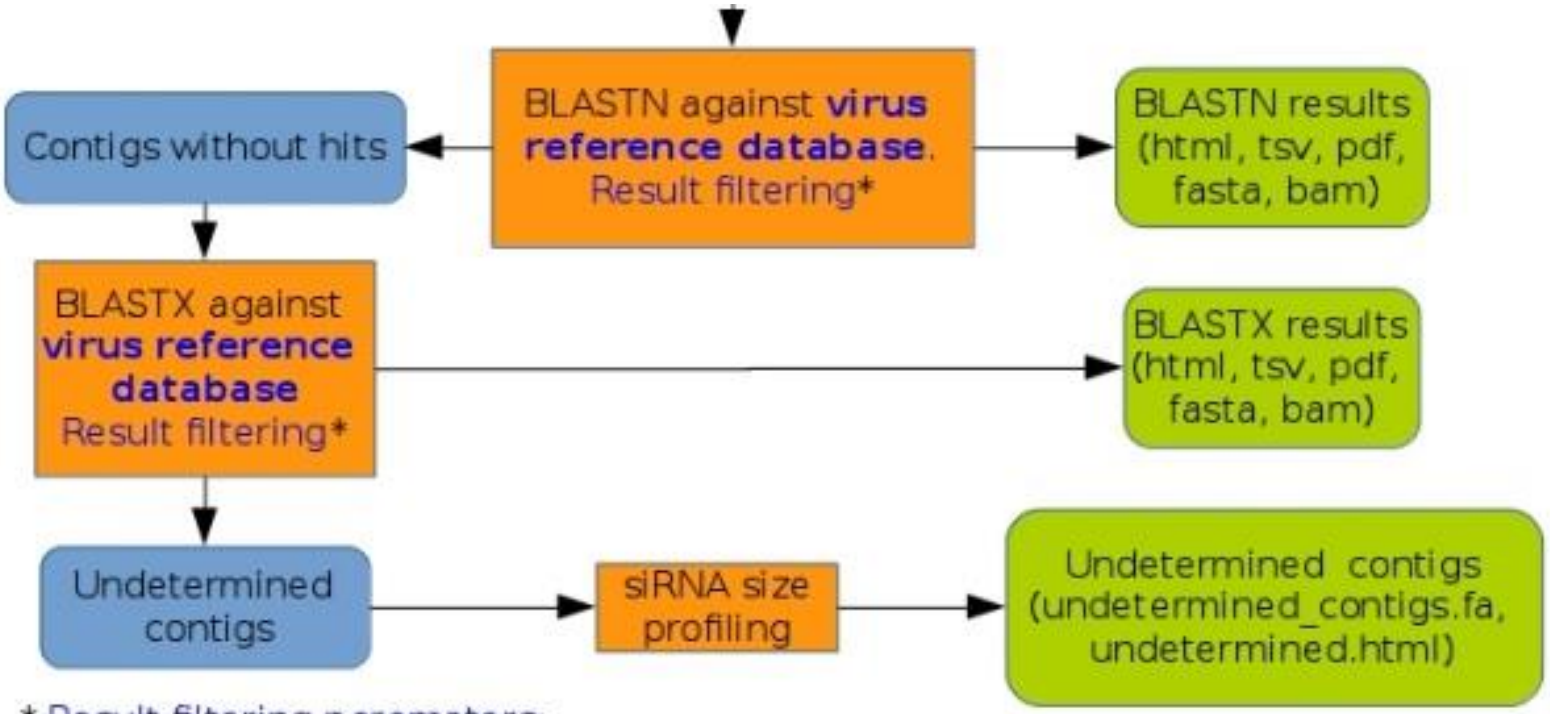
- **Yes, because it improves the Velvet assembly and increases the efficiency of virus detection**
  - enriches virus-derived siRNAs, reduces noise from host-derived small RNAs
- **Yes, because some host genomes contain integrated viral sequences**
  - related to extant replicating viruses but are mostly inactive fragments, and could be falsely identified as an infecting virus
- **No, because host genome can have inadvertent viral sequence contamination**

CSC

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



* 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)

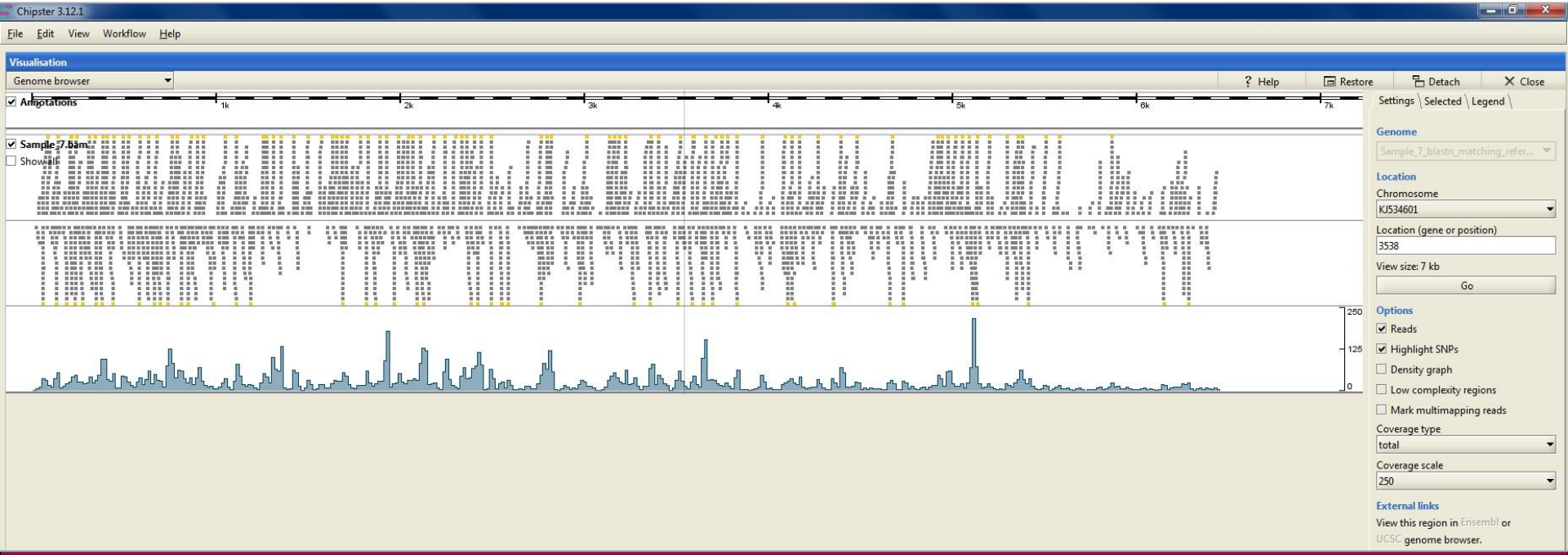
CSC

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

CSC

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
[11/01/17 07:16:27] Align reads to host (Solanum_tuberosum.SolTub_3.0) reference sequences
   30315 reads aligned
[11/01/17 07:16:31] De novo assembly
   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
[11/01/17 07:17:22] Finished
#########################################################################
```

CSC

Vd.log continued

```
-----------------------------------------------------
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.
```

CSC

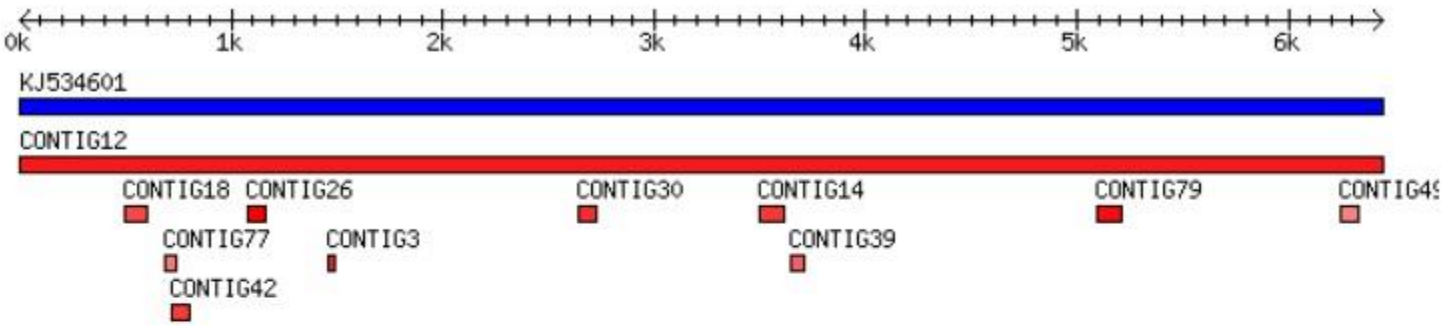
blastn_matching_references.html

Reference	Length	Coverage (%)	#contig	Depth	Depth (Norm)	% Identity	% Iden Max	% Iden Min	Genus	Description
M72416	7568	6710 (88.7)	44	13.1	262.9	97.47	100	90.77	potexvirus	Potato virus X complete genome.
KJ534601	6445	6264 (97.2)	22	27.5	549.3	97.99	100	92.15	potexvirus	Potato virus X isolate SA-CIP, complete genome.

- 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**

CSC

Virus.bn.pdf



Order	Query ID	Query Start	Query End	Subjct Start	Subjct End	Identity	E value	Strand
1	CONTIG3	1	34	1454	1487	33/34(97%)	7e-10	1

Alignment:

```
Query: 1    tgctggactgcttcacaaggatgccagcttatgc 34
            |||||||||||||||||||||||||||| |||||
Sbjct: 1454 tgctggactgcttcacaaggatgccagcctatgc 1487
```

- Blue = virus sequence
- Red = contigs (lighter color indicates lower identity)

CSC

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

Showing 37 rows of 37 and all 14 columns

#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
CONTIG12	GAAAACTAAA...	6445	KJ534601	6445	potexvirus	Potato virus X isolate SA-CIP, complete genome.	1	6445	1	6445	6332/6447(98%)	0.0	1
CONTIG32	TTTCGAAAACT...	5308	M72416	7568	potexvirus	Potato virus X complete genome.	5	5308	1	5307	5119/5309(96%)	0.0	1
CONTIG56	CATTACCTTCC...	1397	M72416	7568	potexvirus	Potato virus X complete genome.	2	1396	5234	6628	1361/1395(97%)	0.0	1
CONTIG56	CATTACCTTCC...	1397	M72416	7568	potexvirus	Potato virus X complete genome.	71	1203	6436	7568	1104/1134(97%)	0.0	1
CONTIG68	GCCTTTGTGAA...	209	M72416	7568	potexvirus	Potato virus X complete genome.	1	201	6235	6435	197/201(98%)	1e-101	1
CONTIG68	GCCTTTGTGAA...	209	M72416	7568	potexvirus	Potato virus X complete genome.	1	201	7368	7568	197/201(98%)	1e-101	1
CONTIG20	CAAAAGAAAG...	163	M72416	7568	potexvirus	Potato virus X complete genome.	1	163	4165	4327	157/163(96%)	3e-74	1
CONTIG79	CAGTCCACCT...	149	KJ534601	6445	potexvirus	Potato virus X isolate SA-CIP, complete genome.	1	114	5098	5211	113/114(99%)	4e-58	1
CONTIG79	CAGTCCACCT...	149	AF172259	6435	potexvirus	Potato virus X complete genome.	1	149	5098	5246	140/150(93%)	2e-56	1
CONTIG14	CAAATGCAAC...	128	KJ534601	6445	potexvirus	Potato virus X isolate SA-CIP, complete genome.	1	120	3495	3614	115/120(95%)	5e-51	1

- 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

CSC

undetermined.html

Undetermined contigs

* potential novel virus contigs are highlighted in green

Contig		siRNA size distribution																	
ID	Length	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	21-22 (%)	Depth	Depth (Norm)
CONTIG15	184	0	0	0	502	618	13	18	0	0	0	0	0	0	0	0	0	97.31	135.15	54.06
CONTIG62	171	0	0	0	702	194	15	13	0	0	0	0	0	0	0	0	0	96.97	115.01	46.00
CONTIG31	137	0	0	0	1722	606	22	12	0	0	0	0	0	0	0	0	0	98.56	366.15	146.46
CONTIG25	129	0	0	0	874	330	18	11	0	0	0	0	0	0	0	0	0	97.65	203.30	81.32
CONTIG74	121	0	0	0	1643	320	11	8	0	0	0	0	0	0	0	0	0	99.04	347.02	138.81
CONTIG66	113	0	0	0	447	175	14	5	0	0	0	0	0	0	0	0	0	97.04	120.88	48.35
CONTIG27	90	0	0	0	1507	297	65	7	0	0	0	0	0	0	0	0	0	96.16	441.59	176.64
CONTIG59	87	0	0	0	85	32	8	19	0	0	0	0	0	0	0	0	0	81.25	35.97	14.39
CONTIG9	82	0	0	0	7	9	16	80	0	0	0	0	0	0	0	0	0	14.29	31.89	12.76
CONTIG45	78	0	0	0	409	127	6	2	0	0	0	0	0	0	0	0	0	98.53	148.38	59.35

- Green color indicates potential virus contigs (21-22nt >50 %)

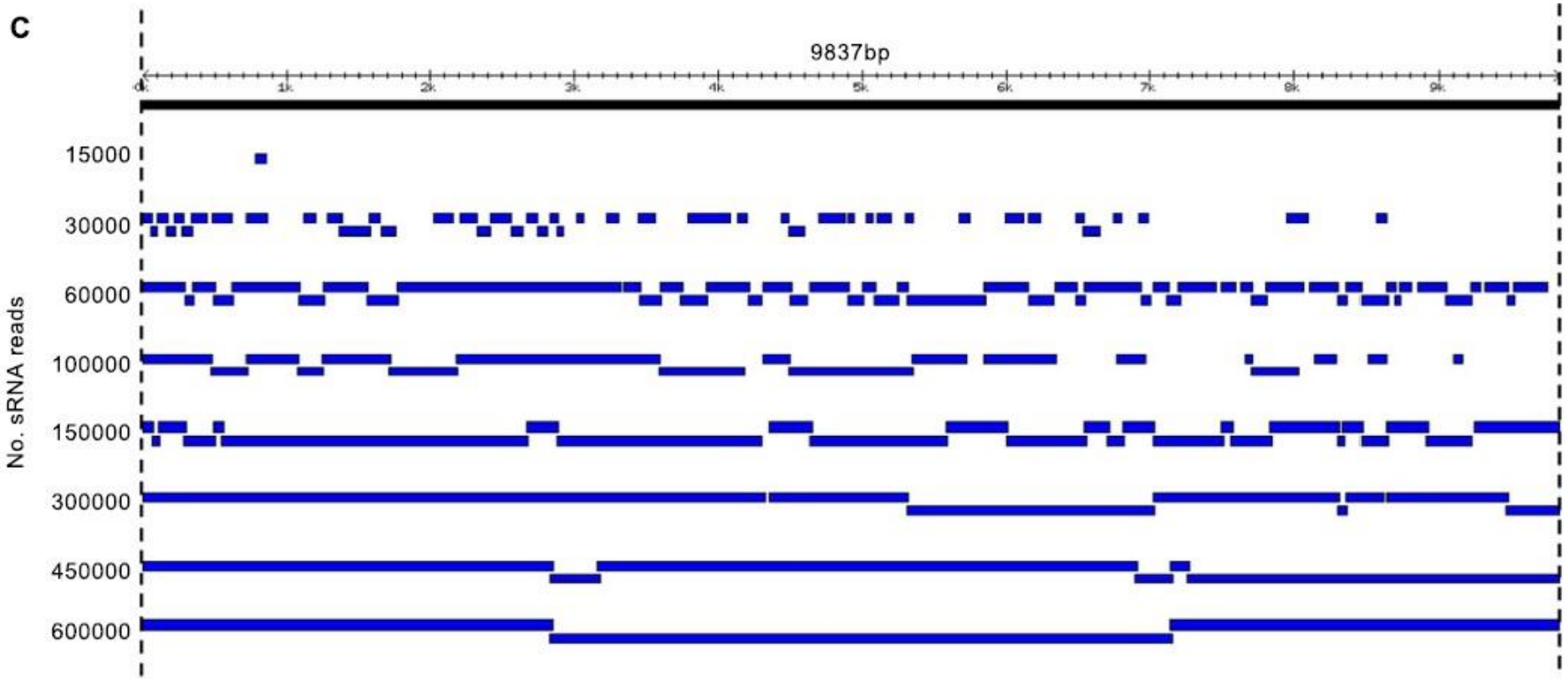
CSC

VirusDetect sequence files and BAM

- **virusdetect_contigs.fa**
  - Sequences of non-redundant contigs derived through reference-guided and *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.

CSC

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

CSC

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