Back to basics with VirDisc: integrating High Throughput Sequencing in the framework of plant virus diagnostics

M. Botermans*, M. Westenberg, M. Visser, L.P. van der Gouw, P.P.M. de Koning, J.W. Roenhorst, B.T.L.H. van de Vossenberg
Netherlands Institute for Vectors, Invasive plants and Plant health (NIVIP), National Plant Protection Organization (NPPO), Netherlands Food and Product Safety Authority (NVWA), Wageningen, Netherlands
Corresponding author: m.botermans@nvwa.nl

S-file 1. Guide to Analyse and Interpret Output Data: Methodology and Sequential Steps

Table 1 gives an overview of essential output of the VirDisc bio-informatic pipeline and how to interpret the data. The alien control sample is assessed first followed by the diagnostic samples. Data generated in the same sequencing batch are analysed and reported batch-wise to identify possible contamination between samples.

Table 1. Interpretation of output data

Sample	Output	File type	What is it?	What to do with it?
Alien control sample	All output	various	Essential output and output for troubleshooting	Ensure that the bio-informatic pipeline functioned as expected and all essential output and output for troubleshooting is generated (see 2.5.4)
	_rRNA_ERCC_statistics	table	Library prep control for rRNA depletion efficiency Library prep control	1. Document mapped reads of rRNA and determine number of rRNA reads in dataset (pass/fail criterion: $reads_{total} - reads_{rRNA} \geq 1.8$ Gb) 2. Assess mean read depth for ERCCs -00019, -00053, -00078 and 00084 (pass/fail criterion: ≥10x)
	_Plant_*rbcL*_Krona_control.html	Html file	BLAST output for the host identity control based on cpDNA *rbcL* sequences	1. Verify taxonomic identity of the host (pass/fail criterion: Genus level identity = *Nicotiana*)
Diagnostic sample	_rRNA_ERCC_statistics	table	Library prep control for rRNA depletion efficiency Contamination control	1. Document mapped reads of rRNA and determine number of rRNA reads in dataset (pass/fail criterion: $reads_{total} - reads_{rRNA} \geq 1.8$ Gb) 2. Assess number of mapped reads to ERCC-0130 (pass/fail criterion: <10 reads)
	_Plant_*rbcL*_Krona_control.html	Html file	BLAST output for the host identity control based on cpDNA *rbcL* sequences	1. Verify taxonomic identity of the host (pass/fail criterion: Family or genus level identity as expected based on material sampled and analysed)

Table 1. Interpretation of output data (continued)

Sample	Output	File type	What is it?	What to do with it?
Diagnostic sample (continued)	_Virus-Viroid_Krona.html	Html file	BLAST (=BLASTn) and DIAMOND (=BLASTx) output of non-plant contigs in the metagenomic assembly from the analysis with all data and the analysis with 1% of all data (= sampled)	1. Document putative viral and viroid species from the nonfiltered_BLASTn and BLASTx output. These represent the contigs assembled with all data. 2. Document putative viral and viroid species from the sampled_nonfiltered_BLASTn and sampled_BLASTx output. These represent the contigs assembled with 1% of all data. 3. Identify all contigs of relevant putative viral and viroid species in relation to the diagnostic question for further analysis by selecting the corresponding section in the Krona chart.
	_Pfam results	table	Overview of Pfam domains identified in the metagenomic assembly (all data and sampled)	1. Identify contigs with putative viral Pfam domains by filtering the output using for instance “vir”, “coat” and “movement”. Retrovirus domains can be ignored as these are typically incorporated in the host genome. 2. Cross check contigs with putative viral Pfam domains with all putative viral and viroid contigs selected from _Virus-Viroid_Krona.html to identify putative viral contigs not represented in the current sequence database.
	_(sampled) assembly	Assembly	*De novo* assembly with read mapping generated from all data and sampled data	1. Assess the quality of the *de novo* assembly for the putative viral and viroid sequences selected for further analysis, for instance: a. quality of read coverage: forward and reverse oriented paired reads b. dips in read coverage resulting in unreliable consensus sequences c. nonspecific mapping of single reads d. presence of multiple genotypes represented by partial nucleotide polymorphisms e. incorrect scaffolding of contigs producing a chimeric consensus sequence 2. Document sequence length and mean read depth for each contig of the selected putative viral and viroid species, taking into account the expected number of genomic sequences and expected sequence length for that particular species. ICTV (https://ictv.global) and ViralZone (https://viralzone.expasy.org) can be used to determine the expected number of genomic fragments.
	(sampled)_chunked	multifasta	Consensus sequences extracted from the *de novo* assembly for all data and sampled data	1. Create a new sequence list of consensus sequences of selected putative viral and viroid sequences that are deemed reliable for further analysis based on the visual assessment of the *de novo* assembly.

Controls

For a reliable test result to be obtained, the external alien control is included for each series of nucleic acid extraction, sequencing and analysis of VirDisc data. In the VirDisc test, the alien control sample is used to monitor both positive and negative control steps. In addition, the bio-informatic pipeline produces output for diagnostic samples that are used as positive and negative control steps. The alien control sample consists of *Nicotiana benthamiana* RNA extract (20 ng/µL) spiked with ERCC RNA Spike-In Mix (Thermo Fischer Scientific, cat. 4456740) at a v:v ratio of N. benthamiana: ERCC of 99:1. Analysis of virus and viroid content in the alien control and diagnostic sample, in the context of the library preparation and sequence batch, is used to detect potential contamination.

Negative control steps for the alien control:

- The alien control serves as negative control for virus/viroid content as no viruses or viroids are expected to be present.

Positive control steps for the alien control:

- Low level target detection is monitored with ERCCs -00019, -00053, -00078 and 00084 present in the alien control sample
- The plant *rbcL* gene is used for host species verification
- Control of rRNA depletion efficiency

Negative control steps for the diagnostic samples:

- ERCC-00130, which is present at a high concentration in the alien control, is monitored in datasets generated for the diagnostic samples to identify potential contamination during library preparation and sequencing.

Positive control steps for the diagnostic samples:

- The plant *rbcL* gene is used for host species verification
- Control of rRNA depletion efficiency

Potential contamination control steps

When (near) complete virus and viroid sequences from the same species are obtained in more than one sample from the same sequence batch, several steps can be taken to assess if these are expected to be truly present in the diagnostic sample or if they are the result of contamination.

- The identified viral/viroid species is in agreement with the contextual information (e.g. virus – host combination, epidemiological information)
- Expected mean read depth of the viral/viroid genome in relation to the material sampled
- Sequence similarity and mean read depth differences of virus/viroid genomes among the different samples

Interpretation of results:

Verification of the controls for the alien control sample

- ERCCs -00019, -00053, -00078 and 00084 present at ≥10x mean read depth
- Free from virus and viroid sequences
- Host species verification with the *rbcL* gene at least at genus level
- $reads_{total} - reads_{rRNA} \geq 1.8$ Gb

Verification of the controls for the diagnostic sample

- ERCC-00130 in datasets of diagnostic samples ≤ 10 reads
- Host species verification with the *rbcL* gene at family or genus level
- $reads_{total} - reads_{rRNA} \geq 1.8$ Gb

When these conditions are met, clustering analysis of the structurally annotated (near) complete virus and viroid sequences is used for (species) identification. Tests should be repeated if any contradictory or unclear results are obtained, or end-users have to justify why the obtained results can still be used for detection and identification of the virus or viroid specie