VirusCurateAU: Minimum requirements and best practice for vouchered viral specimens in plant pathogen reference collections

Agriculture Victoria Research

June/2021

VICTORIA State Government | Jobs, Precincts and Regions

AGRICULTURE VICTORIA

Author: Linda Zheng, Brendan Rodoni and Fiona Constable

Project RDC Number: 4-EM51J3A

Project CMI Number: 6410

Department of Jobs, Precincts and Regions
1 Spring Street Melbourne Victoria 3000
Telephone (03) 9651 9999

Contents

The baseline requirements of VirusCurateAU

VirusCurateAU is intended to be a network that connects government (APPD contributors totalling 18 local databases currently) and private plant virus collections that hold both curated and noncurated plant virus specimens.
The baseline requirements are:

1. All APPD contributors that hold plant virus isolate collections become a part of the network
2. All members of the VirusCurateAU network ensure that they identify key virus isolates of significance to their region and to Australia and curate them to the minimum standard outlined in this document.
3. All local collections/databases for VirusCurateAU require:
    - The identification of the physical specimen is in-line with the species demarcation criteria listed by the ICTV (https://talk.ictvonline.org/ictv-reports/ictv_online_report/).
    - Meeting the minimum requirement of the metadata associated with a specimen/record/accession as identified in this document (Table 1), including images of the specimen and associated signature molecular data.
    - All vouchered specimens designated to be a curated voucher specimen need to be flagged in the local databases with level of completeness (i.e. curated for specimens meeting all minimum requirements and working isolates for specimens meeting most requirements).
    - Validation of the curated records before uploading to the centralised hub.
    - On-going support for the storage and curation of the physical specimens within each collection.
4. The centralised hub for VirusCurateAU requires:
    - Data availability with easy access to pre-existing databases for data sharing.
    - Ability to extract, transform and load data across multiple platforms (i.e. different local databases).
    - Tiers of data access restrictions for end users.
    - Multi-level data security to ensure data integrity
    - Scalability with regards to users and data.
    - User-friendly interface (i.e. customised web application).
    - Dedicated curator of the central hub, with a high-level understanding of virus taxonomy and demarcation criteria, to double-check quality of the records uploaded, specifically the curated voucher specimens.
    - Dedicated IT personnel for on-going maintenance and data management.

Although current APPD contributors are the primary providers of datasets and are required to be members of the VirusCurateAU network it is strongly recommended that holders of private collections become a part of the network, by linking their databased or submitted curated specimens via the APPD contributors, to improve completeness of plant virus records for Australia.

Minimum requirements for a vouchered specimen including the quality of the physical specimen (in-line with international standards)

The physical standards of a well curated plant virus specimen and the best practice preservation and storage method are:

Quantity and quality

- The original field sample in which a virus was first detected is preferred
- Adequate quantity (preferably 5-10 g fresh weight)
- If possible, collect all parts of the plant, including roots, stems, shoots, flowers and/or fruits
    - these materials can be used to identify the plant host
    - viruses may accumulate at higher concentration in some plant parts
- Material should be free of dirt/soil, insects and/or mould
- Dry material as much as possible when collecting (i.e. remove access moisture and avoid storing in sealed plastic bags)

Preservation and storage

- Store in at least two separate containers
- It may be useful to press a specimen with specific symptoms, however high-quality photographs can be substituted

- Desiccated materials:
    - Store desiccated tissues at -20°C (can be freeze-dried, dried on calcium chloride or silica gel)
    - Dried material can be crushed/ground to help reduce the volume and maximise storage space
- Non-desiccated materials:
    - Fresh plant tissues should be stored at -80°C or liquid nitrogen to reduce degradation
    - It is advisable to snap freeze fresh samples in liquid nitrogen prior to storage at -80°C or liquid nitrogen
- Derivatives
    - Derivatives include homogenates of plant tissues in buffer, purified virus particles or nucleic acid/extracts
    - Homogenates and nucleic acid extracts can be stored at -80°C or stabilized on FTA cards (Cardona-Ospina et al 2019).
    - It may be possible to store virus particles in protic ionic liquids (Byrne et al 2012), but more research is required to determine the effectiveness for different virus species

Maintaining live cultures could be considered for important virus isolates. However, those that are passaged over several years in herbaceous hosts can drift genetically (Pretorius et al 2016). If necessary, maintain live cultures in a secure insect free location, preferably in its original host. Re-test regularly to ensure the continuing presence and genetic stability of the virus.

Plant virus cryopreservation in living tissues (Wang et al 2018) could be useful for long-term preservation of plant viruses that invade meristematic tissues because those tissues can be revived after long-term storage and cryopreservation can ensure genetic stability of the virus isolate.

All specimens should be labelled with a unique identifier that can be linked back to a database containing metadata about the specimen.

The minimum requirements for the vouchered specimen metadata including specimen information, genomic data, images and molecular signatures

The main reference used was a summary for each class of quarantine organisms of the minimum quality standards The minimum quality standards recommended for Australian virus reference collections are listed in table 2.

Table 1 Summary of the recommended minimum requirements for metadata associated with a vouchered plant virus/viroid specimen in VirusCurateAU.

Metadata requirement	Specific information to be held/standard operating procedures and competences required
Data to be stored on each accession	Host plant or other source/substrate from which it was collected
	Date (at least year) of sampling (where available)
	Sampler/collector
	Location collected
	Original specimen number or name given by collector (where available)
	Unique accession number in the collection
	Date of deposit in collection
	Preservation conditions and date preserved
	Reference to accession numbers for duplicates in other collections (where available)
	Depositor (where known)
	Species Type (reference strain) strain (yes or no)
	Date of last viability test
	Images of the accession (symptoms)
	Sample type (i.e., tissue/fruit/seed/root) - if plant tissue not available, nucleic acid

	extracts are accepted
	Physical quantity available in collection
	Link to publications/references pertaining to accession where available
	Voucher specimen flag (tiered system, if any information is missing, then flag it as working isolate)
Identification methods	Molecular identification method/s used for identification (i.e. DNA/RNA/barcoding/HTS)
	Sequence data used for identification (need to adhere to ICTV recommendations on virus species demarcation)
	General symptom description (if no symptoms were observed, note down symptomless)
Storage facilities	Location of specimen
	Contact details for persons responsible for the collection
Purity	Purity of accession and methods of checking
Chain of accession	Record keeping for movement of accessions in and out of the collection
Data sharing and public access	Level of access to associated molecular data (tiered system)
	Customer data
	Sharing procedures for selected data with another entity/database

An additional "phenotype" for identification could be an electron micrograph of the associated virus particles. Additional phenotype data could be information about experimental host range and the symptoms observed after inoculation of a virus isolate.

Virus taxonomy is driven by nucleic acid sequence data and the sequence used for curation of a virus isolate must align with International Committee on Taxonomy of Viruses (ICTV) guidelines. Therefore, the minimum requirement for a curated virus specimen is the gene sequence used for virus genus and species demarcation as identified by the ICTV (https://talk.ictvonline.org/). Demarcation criteria differ between virus families for example viruses in the genus *Betaflexiviridae* are primarily classified based on their replicase gene and for borderline cases the coat protein gene sequence may be used, but *Potyviridae* species are demarcated based on the entire polyprotein open reading frame, which is a near whole genome sequence. Comparisons should be made to the exemplar species (https://talk.ictvonline.org/taxonomy/vmr/) for accurate assignment to a species, but phylogenetic comparisons to other strains of the same species may be useful.

Best practice for the characterisation of viral isolates

For best practice characterisation of viral isolates:

- the natural (original) host must be identified, and through the description of plant habit and parts, such as leaves, stems, flowers, roots and fruits. Where these are not available plant DNA barcodes may assist with identification.
- the associated symptoms must be recorded.

Biological properties are highly desirable for virus characterization and include:

- Particle morphology
- Host range and description of associated symptoms
- Mode of transmission – including seed, pollen, mechanical and vectors
- Confirmation of association to disease though isolation and back inoculation to the original host species.

Genetic data of a virus isolate must be included in its characterization, as virus taxonomy is now largely reliant on genomic sequences. As a minimum, the genetic data required for determination of a virus species must be in accordance with the virus species demarcation criteria used by the ICTV, which can differ from family to family. Ideally a full genome or at least the full-length sequence of the coding regions should be used because they are more taxonomically informative, i.e. for the identification of recombination events, which may contribute to virus evolution. The following are the recommended genetic information required for a curated virus specimen

- Full genome or at least that of the coding regions of a virus specimen.
    - When obtained by metagenomic high throughput sequencing, the genome arrangement is preferably confirmed by Sanger sequencing of overlapping PCR amplicons. Sanger sequencing is critically important when there is low sequence depth (i.e., low number of viral sequence reads across the genome or parts of the genome), and the genome has gaps or is incomplete.
    - When derived from PCR and Sanger sequencing of the amplicon – sequences generated by PCR should overlap by a minimum of 100 nucleotide bases to ensure sequence integrity. Cloning and Sanger sequencing the insert at least three times in each direction can resolve ambiguous nucleotides, particularly in the ends near the primers: it is not necessary, but it is desirable.
- Many viruses have multipartite genomes of which each segment packages separately – it is useful to have all segments
- Metagenomic sequencing can identify the virome of a sample, which may assist in elucidating cause of virus associated symptoms.

The development of standard operating procedures for the characterisation of virus specimens

Characterisation of a virus specimen must ensure it meets the minimum standard for a curation:

1. Accurately identify the plant host. If uncertain one of two methods can be used:
    a. Morphology/botanical key
    b. Barcode
2. Describe the geographical location of the host
3. Provide the collection date – at least the year but the exact date is referable
4. Describe if the specimen is plant tissues, purified virus or nucleic acid
    a. Describe the tissue type of the specimen (i.e., leaf, flower, stem, roots, fruit, seed etc)
5. Describe the symptoms observed on the host tissues (the plant may be symptomless) – photographs are preferred.
6. HTS is the preferred method for genetic characterisation of a plant virus, however genome assembly is influenced by several factors (Kutnjak et al 2021, Massart et al 2019):
    a. Depth and quality of HTS data can be influenced by sample quality and virus titre
    b. Sample quality and virus titre may influence the nucleic acid extraction used, the use of ribosomal RNA depletion for enrichment and the sequencing depth required for HTS
    c. The sequencing platform that it used (i.e., short and long read platforms),
    d. Sequence (read) quality and depth must be sufficient to assemble the coding regions of taxonomic importance according to the ICTV and to account for recombination events
    e. The bioinformatic pipeline chosen for genome assembly
7. If HTS is not available, Sanger sequencing of the PCR amplicon of a taxonomically informative gene is needed. Where possible the entire coding region is required (https://talk.ictvonline.org/). Appendix E outlines the standard operating procedure in greater detail.

1. PLANT TISSUE SAMPLING AND COLLECTION

1.1 Sample collection

When sampling, symptomatic, and surrounding plant tissue should always be collected. If no symptoms are showing, fully expanded young leaves that are not showing signs of senescence are optimal. Leaves from different part of the same plant should be sampled, also collect tissue from different part of the plant including stems, seeds and roots.

Plant material should be collected dry and free of dirt/soil, insects and/or mold. Prolonged wetness promotes the growth of fungi and bacteria that contribute to a shortened storage life and degradation of samples so always keep plant samples as dry as possible after collection.

Leaves - a fresh weight of minimum 5g should be collected for leaves and placed into a sterile container/bag with the air expressed prior to sealing. If the materials are excessively wet, they should be shaken prior to bagging to remove excess water. Blot the leaves dry with absorbent paper (i.e. newspaper) prior to placing them between layers of absorbent paper (do not use tissue paper, because this can disintegrate when wet and become difficult to remove from the sample). Press and dry leaf material, making sure to spread out the leaves out so they do not overlap. If the leaves are particularly fleshy, change the absorbent paper daily until the leaves dry out. If tissue is dusty or covered in sooty mold or scale insects, swab them with water or alcohol to clean them.

Stems - the sections of stems collected should have both healthy and diseased areas of tissue. The tissue should be wrapped in absorbent paper such as newspaper individually to avoid damages.

Fruit - choose samples in the early to intermediate stages of symptom development. Do not wrap the sample in plastic, instead wrap them separately in dry absorbent paper (i.e., newspaper).

Seeds – seeds can be stored in paper bags.

Roots - shake off excess soil or, if possible, wash the roots gently without scrubbing to avoid loss of root tissue. Soil contains many microorganisms that readily colonise dead or dying tissue and can interfere with the recovery of pathogens from diseased tissue. Wrap the roots in absorbent paper and store in paper bags when dried or in paper bags for transport back to the laboratory. For herbaceous plants, approximately 25–100 g of root tissue should be sufficient (the lower weight is suitable for vegetables, while the higher weight is more applicable to plants with large roots, such as banana). For woody plants, it may be necessary to excavate to a depth of up to 30 cm near the base of the tree or until roots are found showing the margin between healthy and diseased tissue.

Samples can be dispatched immediately for testing or refrigerated (4°C) until dispatched.

1.2 Labelling

The collection details that should accompany specimens include:

- name of host plant and part of plant affected
- precise location, town, state, province, district, country (longitude and latitude and altitude, if known). A global positioning system (GPS) is the best means of obtaining precise coordinates. GPS-determined coordinates enable accurate distribution maps to be developed for plant pathogens.
- collection date
- collectors' names (collection number if given)
- disease symptoms and severity (eg number of plants affected).

All specimens submitted to an herbarium should be properly labelled. The name of the person submitting the specimen and their contact details are essential.

1.3 Photography

Photographs of plant disease symptoms are extremely useful, because once the plant tissues are processed for long-term storage or stored for a long time, the distinguishable symptoms might fade. Sometimes the plant tissue might be processed into smaller sections and/or ground to a powder for long term storage, which means all observable symptoms will be lost. As a requirement for a vouchered plant virus/viroid specimen, high-quality photographs of the plant material (ideally with symptoms) must be taken.

If possible, digital image of the plant disease specimen should be taken while collecting in the field as these will allow you to build up a pictorial reference collection of plant disease symptoms as they occur in nature.

Photographing plant disease specimens in the laboratory allows greater control over environmental conditions, although the conditions of samples might not be optimal if they were collected and stored improperly. Light grey is the best background for photographs; black and white backgrounds can result in the under or over exposure of the subject.

1.4 Storage

The plant material must be stored in at least two separate containers. It may be useful to press a specimen with specific symptoms, however high-quality photographs can be substituted.

Specimens can be stored either desiccated or non-desiccated materials.

Plant material can be desiccated via freeze-drying or with desiccants such as calcium chloride or silica gel. Once dried, samples can be stored in sterile glass ampoules or vials or plastic containers and stored at 20°C for a prolonged period. Dried material can be crushed/ground to help reduce the volume and maximise storage space.

For non-desiccated materials, fresh plant tissues should be stored at -80°C or liquid nitrogen to reduce degradation. It is advisable to snap freeze fresh samples in liquid nitrogen prior to storage at -80°C or liquid nitrogen.

Derivatives include homogenates of plant tissues in buffer, purified virus particles or nucleic acid/extracts. Homogenates and nucleic acid extracts can be stored at -80°C, or stabilized on FTA cards (da Cunha Santos et al., 2018).

Live cultures could be considered for important virus isolates. However, those that are passaged over several years in herbaceous hosts can drift genetically (Pretorius et al., 2016). If necessary, maintain live cultures in a secure insect free location, preferably in its original host. Re-test regularly to ensure the continuing presence and genetic stability of the virus.

All specimens should be labelled with a unique identifier that can be linked back to a database containing metadata about the specimen.

2. VIRUS CHARACTERISATION USING HIGH THROUGHPUT SEQUENCING

The visible symptoms of virus infection are often discernible only to the experienced diagnostician. There are two main types of virus symptoms: those resulting from primary infection of host plant cells such as lesions; and those caused by secondary or systemic infection such as mosaic or mottling. Symptoms of plant viruses vary largely depending on the host/virus species/variety/strains and growing conditions in the field and as such diagnosis based on symptoms alone is not advised and should be used as a guide only.

For a vouchered plant virus/viroid specimen, molecular techniques are required for the identification. This SOP will focus on the use of deep sequencing (RNA-seq) for the characterising of a plant virus/viroid specimen. This is considered a metagenomic sample where the sequence data generated will consists of the virome of the host. Figure D.1. demonstrates the workflow for HTS sequencing and confirmation of viruses that are detected.

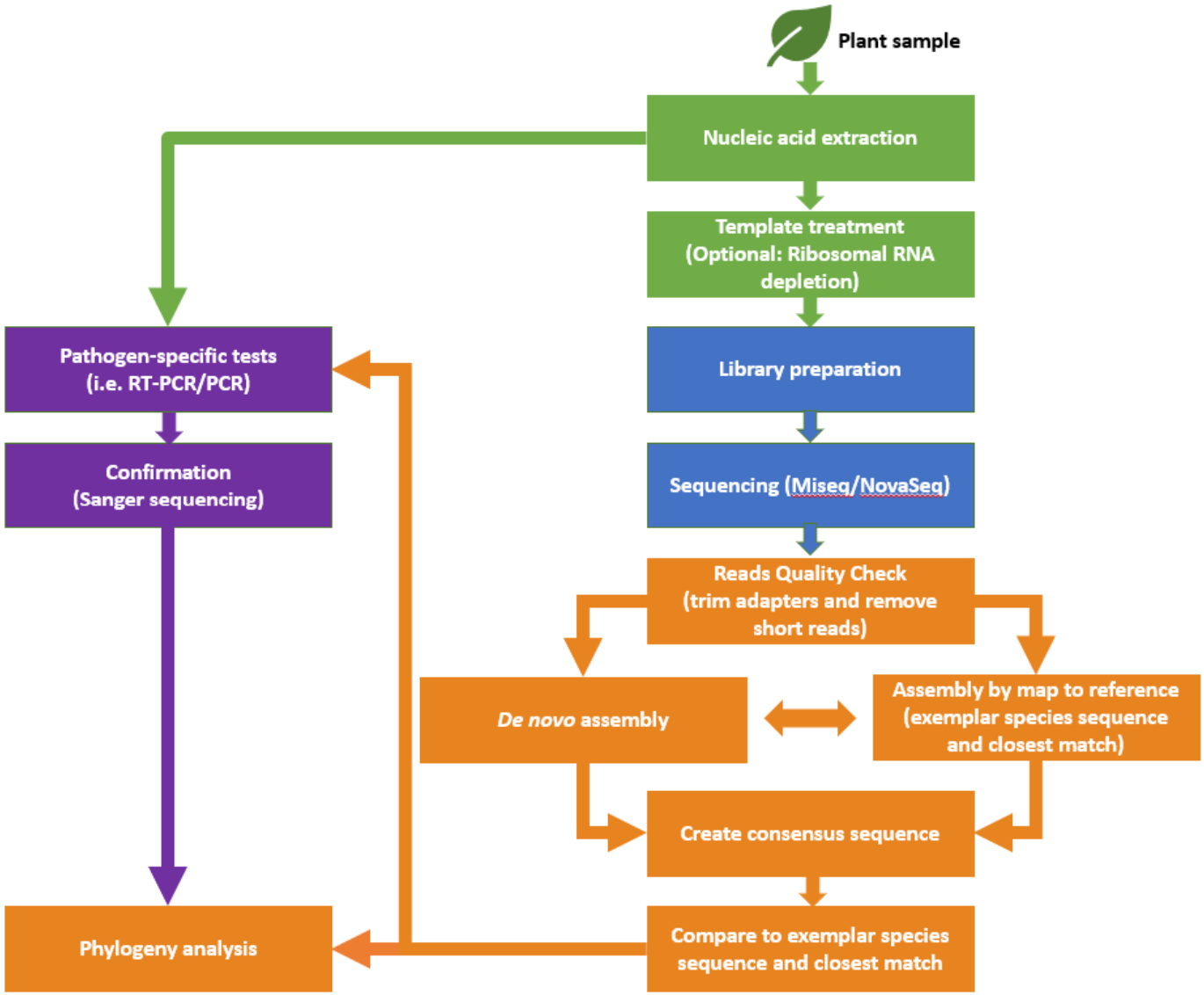

Figure 1 Flow chart for virus detection using high throughput sequencing (HTS)
The different coloured boxes represent different aspects of identifying and confirming a virus in a plant sample by HTS.
Green – sample preparation: Nucleic acid is extracted from a sample and then treated to remove ribosomal RNA to ensure sequence data is enriched for viral RNA.
Blue – HTS sequencing: The RNA is used for library preparation and sequencing using an Illumina platform.
Orange – bioinformatics: During the bioinformatics process the sequence data is trimmed of adapters and short reads are removed to ensure quality. An initial *de novo* assembly is done to create contiguous sequences (contigs) from the reads. If a specific virus is suspected the reads can be assembled to a reference sequence (mapped) – this might be process might be driven by an initial test result using another method or from the results of the *de novo* assembly. Either assembly process will result in a consensus sequence of a virus target, which should then be compared to the sequence of the exemplar strain of the virus and to the closest matching strain, if it is not the exemplar strain. A phylogeny can be done, if required, to compare the sequence to all strains of a virus or to the members within a genus and family if a novel virus species is detected. Phylogenies should be done with the genome regions used by the International Committee on Taxonomy of Viruses (ICTV) for species demarcation.
Purple – Confirmation of detection: Confirmatory testing should be done using PCR methods and Sanger sequencing of the PCR amplicons. If the virus is significantly divergent or novel, the genome arrangements of the virus should be confirmed by generating overlapping (at least 100bp) PCR amplicons and Sanger sequencing.

2.1 RNA Extraction

The method given here utilizes the RNeasy® Plant Mini Kit. Other kits that produce high quality RNA of sufficient concentration can also be used.

Equipment

1. 2-20 µL, 20-200 µL and 200-1000 µL micropipettes and sterile filter tips
2. Autoclave
3. Autoclave bags
4. Balance (at least 2 decimal places)
5. Disposable gloves
6. Microcentrifuge
7. Sterile microcentrifuge tubes
8. Paper towel
9. RNeasy® Plant Mini Kit (Qiagen™)
10. Sharps container
11. Sterile scalpel blades and scalpel blade handle
12. Water bath or heat block set at 70°C
13. Weighing boats
14. Vortex mixer
15. Plastic disposable Pasteur pipettes
16. Homex tissue macerator and Homex bags (Bioreba AG/BioSys) or an autoclaved mortar and pestle

Reagents

Guanidine thiocyanate extraction buffer (Table 1) (Mackenzie et al., 1997)

Table 1 Recipe for the Guanidine thiocyanate extraction buffer

Chemical	**Amount**	**Final Concentration**
Guanidine thiocyanate ($CH_5N_3 \cdot CHNS$)	23.64 g	4 M
PVP-40 (polyvinylpyrrolidone)	1.25 g	2.5% (w/v)
3 M Sodium acetate ($C_2H_3NaO_2$)	3.33 mL	0.2 M
0.5 M EDTA ($C_{10}H_{16}N_2O_8$)	2.5 mL	0.025 M
Add sterile DNA/RNAse free water to final volume of 50 mL. Buffer can be stored at room temperature for 3-6 months.		

1. 100% β-mercaptoethanol (C_2H_6OS)
2. 20% N-Lauroylsarcosine solution (w/v)
3. Ethanol (100%)

Method

1. Determine the number of samples and label plastic tubes accordingly.
2. Use new clean gloves and scalpel blades for each sample.
3. Cut each new sample on fresh paper towel on the bench.
4. Weight out ~200 mg of fresh plant tissue (or 100 mg if freeze dried).
5. Place sample in Homex bag/mortar.
6. Add 2.0 mL of MacKenzie buffer.
7. Add 1/100 volume of β-mercaptoethanol (v:v) to each sample in the fume hood.
8. Homogenise in the fume hood.

9. Pipette 1.0 mL of the slurry into a labelled microcentrifuge tube.
10. Add 100 µL of 20% N-Lauroylsarcosine to each tube and vortex to mix.
11. Incubate tubes at 70°C for 15 minutes.
12. Spin tubes in microcentrifuge for 2 minutes at maximum speed (>= 13, 000 rpm).

Take the supernatant without disturbing the pellet formed and continue with step 4 of the "Purification of Total RNA from Plant Cells and tissues and Filamentous Fungi" as per manufacturer's instructions (page 50 of the RNeasy® Mini Handbook, June 2012).

The quality and concentration of the extracts can be determined using the Nanodrop Spectrophotometer (version 1000; ThermoFisher Scientific).

Quantification of nucleic acids using the Qubit Fluoremeter

A more precise measurement of the concentration of the nucleic acids can be made with the Qubit Fluorometer (ThermoFisher Scientific) since the kits can be used to measure DNA and/or RNA only.

Single-stranded (ss) RNA can be measured using the Broad Range Quant-iT™ RNA Assay Kit (cat. Q10213; ThermoFisher Scientific) following the manufacturer's protocols (Figure 1). (https://www.thermofisher.com/order/catalog/product/Q10213?SID=srch-srp-Q10213#/Q10213?SID=srch-srp-Q10213). Please note that links to protocols can change over-time. If the link posted here is no longer working, please visit the manufacturer's website for the specific protocol.

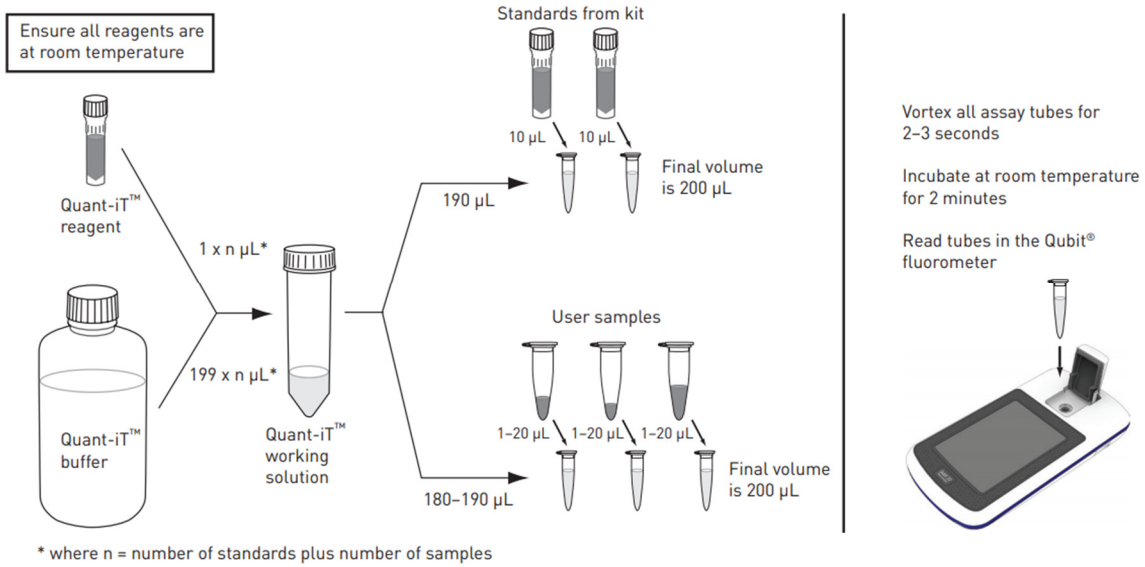

Figure 1 Overview for using the Quant-iT™ RNA BR assay in the Qubit® Fluorometer

2.2 Library Preparation

Different kits from different manufacturers can be used for the preparing of sequencing libraries. However, Ribosomal-RNA removal is highly recommended prior to the preparation of the sequencing library.

2.2.1. RNA-seq with Illumina TruSeq Stranded Total RNA with Ribo-Zero Plant kit

The enzyme used in the TruSeq Stranded Total RNA with Ribo-Zero Plant (cat. 20020610 for 48 samples or cat. 20020611 for 96 samples; Illumina) is the industry gold-standard for ribosomal-RNA (rRNA) removal from plant hosts and highly recommended for RNA-seq of plant viruses/viroids.

To prepare RNA libraries, follow the manufacturer's protocol (https://support.illumina.com//downloads/truseq-stranded-total-rna-with-ribo-zero-plus-rrna-depletion-ref-guide.html).

This kit can be adapted to the Perkin Elmer NEXTflex® Unique Dual Index Barcodes. Changes include replacing the Illumina Adaptor for Illumina for NEXTflex® Unique Dual Index Barcodes 1 - 96 (25 µM).

Important notes:

Half reactions

The libraries can be made with half of the reaction volume recommended by the manufacturer. This will help reduce cost of sequencing for individual libraries.

Input quantity

The input quantity is the extremely important. The recommended input is 0.1 – 1 ug of high-quality total RNA extracted from plant samples, but libraries have been prepared and sequenced successfully for input quantity as low as 0.03 ug without the rRNA-depletion step. This is because rRNA constitutes 70 to 80% of total RNA whereas mRNA is at best only 5%. When the rRNA removal is done, coupled with subsequent clean up, the quantity of the input drops to a point where libraries can no longer be made.

Illumina for NEXTflex® Unique Dual Index Barcodes 1 – 96

When substituting the Illumina adapters with the NEXTflex® Unique Dual Index Barcodes 1 – 96 (Perkin Elmer), use the recommended adapter concentration below (

Table 2).

Table 2 Adapter dilution

Input quantity	Working adapter concentration
100 ng- 500 ng	6.25 µM
50 ng – 99 ng	0.625 µM
< 50 ng	0.31 µM

For rRNA-depleted RNA extracts, please use this (Table 3). Do not perform rRNA-depletion if initial input quantity is lower than 30 ng.

Table 3 Adapter dilution for rRNA-depleted inputs

Input quantity	Working adapter concentration
100 ng- 500 ng	0.625 µM
30 ng – 99 ng	0.31 µM

Fragmentation

The standard fragmentation time in the Illumina TruSeq Stranded Total RNA with Ribo-Zero Plant kit protocol produces fragments with a median size of ~155bp. Adjusting the fragmentation time will either increase or decrease the fragment size.

- With partially degraded RNA or RNA nucleic extracts with low quality scores, decrease the recommended fragmentation time from 8 minutes at 94ºC to 5 minutes avoid over fragmentation of the RNA sample.
- Adjust the fragmentation time to suit your needs. For example, if the library is to be run on a MiSeq Reagent Kit v3 with read length of 2x300bp, then the desired fragment is 300 – 450 bps (Figure 2).

	MiSeq Reagent Kit v2				MiSeq Reagent Kit v3	
Read Length	1 × 36 bp	2 × 25 bp	2 × 150 bp	2 × 250 bp	2 × 75 bp	2 × 300 bp
Total Time*	~4 hrs	~5.5 hrs	~24 hrs	~39 hrs	~21 hrs	~56 hrs
Output	540–610 Mb	750–850 Mb	4.5–5.1 Gb	7.5–8.5 Gb	3.3–3.8 Gb	13.2–15 Gb

	MiSeq Reagent Kit v2 Micro	MiSeq Reagent Kit v2 Nano	
Read Length	2 × 150 bp	2 × 250 bp	2 × 150 bp
Total Time*	~19 hrs	~28 hrs	~17 hrs
Output	1.2 Gb	500 Mb	300 Mb

* Total time includes cluster generation, sequencing, and base calling on a MiSeq System enabled with dual-surface scanning.

Figure 2 Cluster Generation and Sequencing of different kits on the MiSeq system

- If the library is to be run on a NovaSeq 6000 platform with a S4 flow cell (Figure 3), then the ideal fragment sizes should be 150-200 bp.

	NovaSeq 6000 System			
Flow Cell Type	SP	S1	S2	S4
Output Per Flow Cell				
1 × 35 bp	N/A ‡	N/A ‡	N/A ‡	280-350 Gb
2 × 50 bp	65–80 Gb	134–167 Gb	333–417 Gb	N/A ‡
2 × 100 bp	134–167 Gb	266–333 Gb	667–833 Gb	1600–2000 Gb
2 × 150 bp	200–250 Gb	400–500 Gb	1000–1250 Gb	2400–3000 Gb
2 x 250 bp	325-400 Gb	N/A ‡	N/A ‡	N/A ‡
Clusters Passing Filter				
Single Reads	650–800 million	1.3–1.6 billion	3.3–4.1 billion	8–10 billion
Paired-end Reads	1.3–1.6 billion	2.6–3.2 billion	6.6–8.2 billion	16–20 billion
Run Time §				
1 × 35 bp	N/A ‡	N/A ‡	N/A ‡	~14 hr
2 × 50 bp	~13 hr	~13 hr	~16 hr	N/A ‡
2 × 100 bp	~19 hr	~19 hr	~25 hr	~36 hr
2 × 150 bp	~25 hr	~25 hr	~36 hr	~44 hr
2 x 250 bp	~38 hr	~N/A	~N/A	~N/A

* Output and read number specifications based on a single flow cell using Illumina PhiX control library at supported cluster densities. The NovaSeq 6000 System can run 1 or 2 flow cells simultaneously.
§ Run time includes cluster generation, sequencing, and base calling. Run times are based on running 2 flow cells of the same type; starting two different flow cells will impact run time.
Specifications based on Illumina PhiX control library at supported cluster densities.
‡ N/A: not applicable

Figure 3 NovaSeq 6000 System Flow Cell Specifications

2.2.2 Quality check of libraries

Once libraries are prepared, the 2200 TapeStation system (Agilent technologies) can be used for quality checking of the libraries following the manufacturer's protocol (https://www.agilent.com/cs/library/usermanuals/public/G2964-90000_TapeStation_USR_ENU.pdf). A good library should consist of only the desired fragment (between 200 – 300 bps). Any peaks less than 150bp (

Figure 4) indicate the presence of either primer dimers or excess indexes and must be cleaned up (see 2.2.3) before the library can be subjected to sequencing.

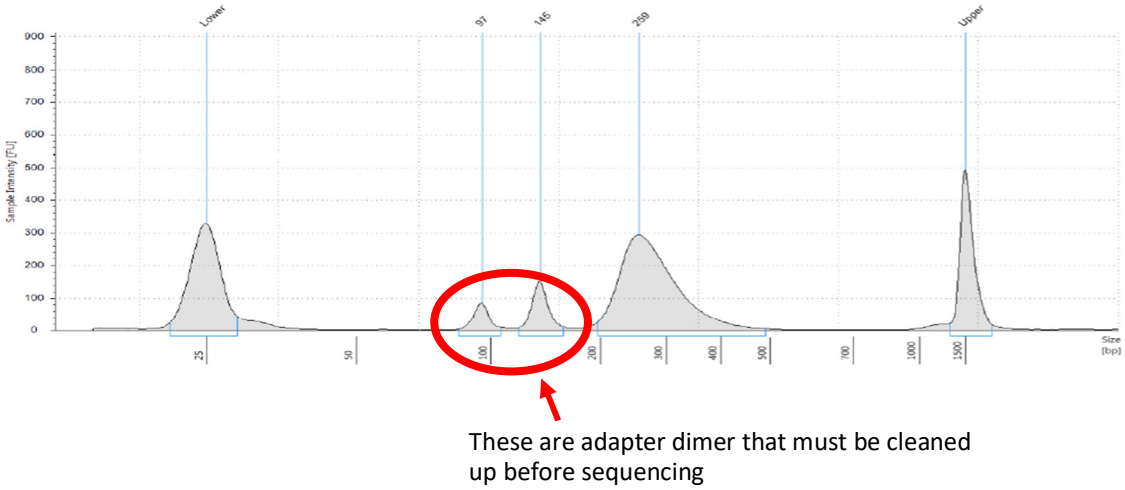

Figure 4 Example of a bad library on a D1000 tape assay on the TapeStation. The peaks between 90– 150 bp (circled) indicates the presence of either primer dimers or excess indexes.

2.2.3 Adapter block

Blocking unligated adapters to prevent index hopping is done following the instructions provided by the manufacturer (https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/reagent/illumina-free-adapter-blocking-reagent-reference-guide-1000000047585-00.pdf).

2.2.4 Pooling of libraries

1. With large number of samples, the individual libraries can be pooled using the Peak Molarity (pmol/L) value determined by the tapestatiom assay. In the example below, the molarity of the library is 10,800 pmol/L which is equivalent to 10.8nM (Figure 4). This value can be used to pool the libraries but it's best that all individual libraries to be pooled based on the TapeStation values be run on the same tape.

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	355	-	21800	-		Lower Marker
398	2780	-	10800	100.00		
1500	250	250	256	-		Upper Marker

Figure 5 An example of a peak table generated by the TapeStation assay

2. Pool individual libraries based on their molarity. I.e. for 1:1 ratio of individual libraries, dilute them to the same concentration and add in equal amounts for the pooled library.
3. Determine the final library concentration using the Qubit Fluorometer and recheck the pooled library using Tapestation. For Illumina sequencing, HSD1000 tape assays are used for DNA concentrations below 30 ng/µL and D1000 tape assays are used for DNA concentrations above 30 ng/µL.

4. Dilute the final pooled library to between 10-15nM using the following formula:

   nM of library = concentration (ng/µL)/(fragment size (bp) x 650) x 1000000
5. Perform Qubit and Tapestation assay (HSD1000 tape) on the diluted pooled library to confirm concentration and size.
   * The quantification of this final library is extremely important as the sequencing platforms are very sensitive to over and/or under-loading.
6. The pooled library is now ready for sequencing.

2.3 Illumina sequencing

The pooled library can be sequenced on the Illumina MiSeq platform or the Illumina NovaSeq 6000 platform.

2.3.1 Using the MiSeq platform

The Illumina MiSeq (MiSeq reagent kit v3) is to be used on the MiSeq platform. Following the manufacturer's instructions for sample preparation for loading, denaturing and diluting libraries (https://www.illumina.com/products/by-type/sequencing-kits/cluster-gen-sequencing-reagents/miseq-reagent-kit-v3.html).

To avoid over and/or under-loading, aim for 10pm when using the Miseq reagent kit V3.

Choose 600 cycles for maximum data output.

2.3.2 Using the NovaSeq platform

The NovaSeq 6000 is the platform of choice for HTS because the cost per gigabase (GB) is much lower than that of MiSeq (~ 4 times lower per GB) and allows a much higher depth of the genome to be sequenced in one run. The NovaSeq 6000 is operated by dedicated technical staff and below is an example of a submission sheet for libraries to be run on one lane of a S4 flow cell.

Please note that the values provided here on the example is only a guide. Different operators will have different requirements for the submission of the libraries, so please adhere to the requirements listed by your operator.

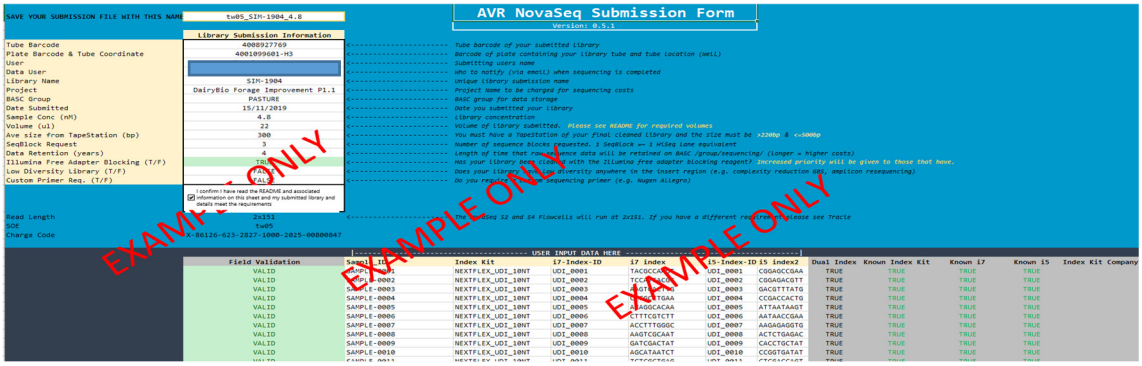

Figure 6 An example submission sheet for libraries to be run one lane on a S4 flow cell on the NovaSeq 6000 platform

3. BIOINFORMATICS WORKFLOW AND PIPELINE

Once the libraries have been sequenced, there will usually be 2 sequence output files per sample. One is "sample name_R1.fastq.gz" and the second one "sample name_R2.fastq.gz" for pair-end sequencing (i.e. 2x150bp). These two files contain the sequence reads generated from each direction of the ligated fragment. The following workflow is used to process these reads (Figure 7).

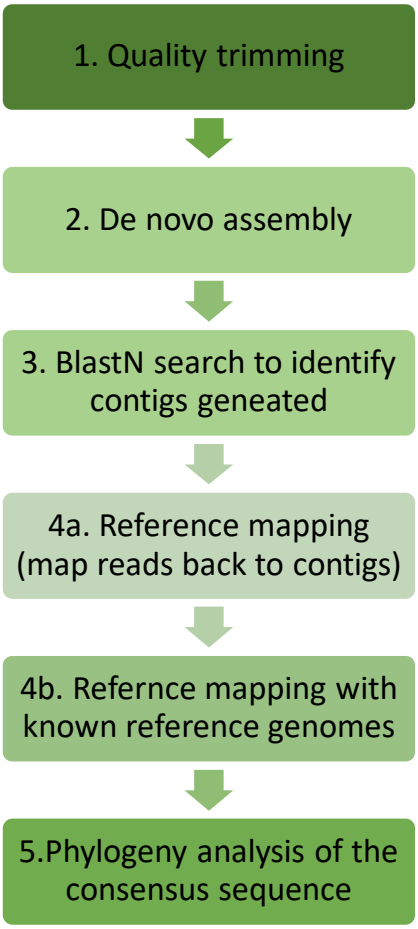

Figure 7 Schematics of the bioinformatics workflow for HTS sequencing processing

Steps 1-4 of the above process can either be done using a super computer cluster dedicated for bioinformatics analysis, commercial softwares (CLC genomics workbench; Qiagen and Geneious Prime; Biomatters) or online servers such as VirDetect (Selitsky et al., 2020) or VirFind (Ho and Tzanetakis, 2014). Phylogeny analysis can be done with any software of choice.

The following process is done using a supercomputer cluster and includes scripts for the specific applications used.

3.1 Read trimming

Before processing, the reads should always be quality checked. This process is usually done using programs that will trim the reads based on the user's input requirements on read length, read quality score and the adapter sequence. The quality trimming will ensure 1) adapters are removed; 2) reads shorter than a specified length are removed. *FastP* (Chen et al., 2018) is used for the example below to trim the sequences with a quality phred score less than 20 and lengths shorter than 50. The script below is for paired end data (gzip compressed)

```
fastp -q 20 -l 50 -i in.R1.fq.gz -I in.R2.fq.gz -o out.R1.fq.gz -O out.R2.fq.gz
```

By default, the HTML report is saved to fastp.html (can be specified with -h option)(Figure 8), and the JSON report is saved to fastp.json (can be specified with -j option).

fastp report

Summary

General

fastp version:	0.20.0 (https://github.com/OpenGene/fastp)
sequencing:	paired end (151 cycles + 151 cycles)
mean length before filtering:	151bp, 151bp
mean length after filtering:	129bp, 129bp
duplication rate:	91.804542%
Insert size peak:	115

Before filtering

total reads:	647.864000 K
total bases:	97.827464 M
Q20 bases:	93.218633 M (95.288817%)
Q30 bases:	88.484755 M (90.449810%)
GC content:	50.205060%

After filtering

total reads:	601.814000 K
total bases:	77.782413 M
Q20 bases:	75.958169 M (97.654683%)
Q30 bases:	72.855551 M (93.665841%)
GC content:	48.529106%

Filtering result

reads passed filters:	601.814000 K (92.892027%)
reads with low quality:	43.438000 K (6.704802%)
reads with too many N:	36 (0.005557%)
reads too short:	2.576000 K (0.397614%)

Figure 8 fastp report in html format

3.2 *De novo* assembly with SPAdes

SPAdes is available to download from GitHub at https://github.com/ablab/spades#sec1.2. SPAdes is also available as a built-in assembler (one of two options) in Geneious Prime.

An example script using SPAdes (Bankevich et al., 2012) is as below, with kmer lengths of 127, 107, 87, 67, 31 (kmer lengths should always be shorter than read length). The output files will be stored in the output folder. The -careful option in SPAdes is to minimize number of mismatches in the final contigs. The -pe1-1 and -pe1-2 are the two files that contain the R1 and R2 reads. Make sure trimmed reads are used.

```
python file_path_of_spades.py \
-k 127,107,87,67,31 \
-o output_folder \
--careful \
--pe1-1 R1.fp.fastq.gz \
--pe1-2 R2.fp.fastq.gz
```

3.3 BlastN search

The final contigs or scaffolds produced by the assembler can now be searched against the NCBI GenBank database using the BlastN (Altschul et al., 1997) function to determine their identities. This can also be done on a dedicated server, a commercial software such as Geneious Primer or directly through the NCBI website (https://blast.ncbi.nlm.nih.gov/Blast.cgi).

The nucleotide databases stored locally in the bioinformatics server was used in the example below for identification.

*Caution – the reliability of this is completely reliant on the accuracy and comprehensiveness of the nucleotide database the contigs are queried against. This means, if the database is not updated regularly then newly published sequences will not be included in the analysis. Given that the GenBank nucleotide database accepts new submission of sequences continuously, the locally stored database WILL ALWAYS be out of date. Also, if the sequence deposited in GenBank is mislabelled, there will be misidentification of the contigs if BLASTn search is the only method used for identification. The best practice is to upgrade the local database (if using) every 3 months or so to ensure the latest viral sequences are being captured.

After blastn is activated, use script below. The script is using the blastn function to query the scaffolds (scaffolds.fasta) against the nucleotide (nt) base with the max_target_seqs of 1 (meaning keep 1 of the aligned sequence) and output the result in format 6 (-outfmt) which is a tabular output with the following columns "qseqid sseqid stitle slen qlen qstart qend mismatch gapopen qcovs qcovhsp pident evalue". The output file is in the .xml format and can be viewed in Microsoft Excel (Figure 9).

```
blastn -query scaffolds.fasta -db /group/blastdb/nt -max_target_seqs 1 -outfmt "6 qseqid sseqid stitle slen qlen qstart qend mismatch gapopen qcovs qcovhsp pident evalue" -out output.xml
```

NODE_1_length_9720_cov_501.121_ID_35155	gi|300068806|dbj|AB570195.1|	Lily mottle virus RNA, complete genome, isolate: ML61	9645	9720	2	5680	75	0	98	58	98.679	0
NODE_2_length_5525_cov_71.6488_ID_14235	gi|1635510314|emb|LR584020.1|	Turnip yellows virus genome assembly, complete genome: monopartit	5719	5525	1787	3167	334	38	25	25	72.811	1.49E-116
NODE_7_length_3850_cov_113.163_ID_10513	gi|916344642|gb|KR107528.1|	Grapevine Cabernet Sauvignon reovirus isolate CS-BR structural protei	3849	3850	2254	2750	122	2	13	13	75.1	1.46E-55
NODE_12_length_3359_cov_581.674_ID_48613	gi|300068806|dbj|AB570195.1|	Lily mottle virus RNA, complete genome, isolate: ML61	9645	3359	1	3359	54	0	100	100	98.392	0
NODE_57_length_1138_cov_649.072_ID_48615	gi|1484033195|gb|MH360242.1|	Lily mottle virus isolate C4 polyprotein (ORF1) gene, complete cds; anc	9626	1138	1	1138	19	0	100	100	98.33	0
NODE_62_length_1120_cov_5098.97_ID_42719	gi|308549671|gb|HM222522.1|	Lily symptomless virus strain LSV-DL, complete genome	8394	1120	355	1068	5	0	95	64	99.3	0
NODE_73_length_1000_cov_4068.98_ID_47183	gi|34915798|emb|AJ516059.1|	Lily symptomless virus complete genome	8394	1000	1	1000	7	0	100	100	99.3	0
NODE_74_length_991_cov_4595.59_ID_48623	gi|1484033214|gb|MH360247.1|	Lily symptomless virus isolate C4 RdRp (ORF1), TGB1 (ORF2), TGB2 (ORF	8355	991	1	991	11	0	100	100	98.89	0
NODE_79_length_936_cov_3457.21_ID_39271	gi|34915798|emb|AJ516059.1|	Lily symptomless virus complete genome	8394	936	1	936	9	0	100	100	99.038	0
NODE_90_length_892_cov_3662.32_ID_47085	gi|34915798|emb|AJ516059.1|	Lily symptomless virus complete genome	8394	892	1	891	8	0	99	99	99.102	0
NODE_91_length_891_cov_3667.1_ID_39937	gi|34915798|emb|AJ516059.1|	Lily symptomless virus complete genome	8394	891	1	891	8	0	100	100	99.102	0
NODE_94_length_872_cov_3751.96_ID_36517	gi|34915798|emb|AJ516059.1|	Lily symptomless virus complete genome	8394	872	1	872	8	0	100	100	99.083	0
NODE_96_length_855_cov_5353.22_ID_33971	gi|308549671|gb|HM222522.1|	Lily symptomless virus strain LSV-DL, complete genome	8394	855	2	855	6	0	99	99	99.297	0
NODE_97_length_855_cov_4712.58_ID_34289	gi|308549671|gb|HM222522.1|	Lily symptomless virus strain LSV-DL, complete genome	8394	855	2	855	8	0	99	99	99.063	0
NODE_105_length_815_cov_2688.8_ID_48617	gi|34915798|emb|AJ516059.1|	Lily symptomless virus complete genome	8394	815	1	815	2	0	100	100	99.755	0
NODE_106_length_808_cov_2184.49_ID_41556	gi|34915798|emb|AJ516059.1|	Lily symptomless virus complete genome	8394	808	1	808	8	0	100	100	99.01	0
NODE_107_length_808_cov_2840.77_ID_41736	gi|34915798|emb|AJ516059.1|	Lily symptomless virus complete genome	8394	808	1	808	7	0	100	100	99.134	0
NODE_108_length_808_cov_2175.41_ID_43425	gi|34915798|emb|AJ516059.1|	Lily symptomless virus complete genome	8394	808	1	808	8	0	100	100	99.01	0
NODE_112_length_803_cov_4181.51_ID_39695	gi|34915798|emb|AJ516059.1|	Lily symptomless virus complete genome	8394	803	1	803	12	0	100	100	98.506	0
NODE_121_length_738_cov_3696.38_ID_46297	gi|47934890|gb|AY620984.1|	Lily symptomless virus isolate lishui02 32 kDa coat protein gene, partia	982	738	2	583	2	0	79	79	99.656	0
NODE_130_length_727_cov_5715.62_ID_33815	gi|308549671|gb|HM222522.1|	Lily symptomless virus strain LSV-DL, complete genome	8394	727	1	726	5	0	99	99	99.311	0
NODE_136_length_710_cov_4126.42_ID_46361	gi|34915798|emb|AJ516059.1|	Lily symptomless virus complete genome	8394	710	4	710	6	0	99	99	99.151	0

Figure 9 An example of a BlastN outout file (output format 6) viewed in excel

3.4 Reference Mapping with Bowtie 2

The mapping tool of choice is Bowtie 2 (Langmead and Salzberg, 2012). Bowtie or bowtie 2 are incorporated into various commercial software such as Geneious Prime (one of two options).

Two types of reference mapping can be done. 1) map the reads to the viral contigs generated by the de novo assembler and 2) map the reads to a reference genome of choice. The following script can be used to do both. When the jobs are done in a server environment, there is a need to copy files to the server and then removing them afterwards to avoid clustering. If the job is run directly, there is no need to copy and remove files after.

```
# Copy read1 and read2 fastq
cp R1.fp.fastq.gz .
cp R2.fp.fastq.gz .

gunzip *.gz

##### Copy all references needed for the job #########
cp reference_genome.fasta .
```

```
#### make the index ####
bowtie2-build reference_genome.fasta reference_genome

####### run the different scripts ########
bowtie2 -p 4 --very-sensitive-local -x reference_genome -1 R1.fp.fastq -2 R2.fp.fastq -S output.sam
samtools view -bS -F 4 output.sam > output.bam
samtools sort -O bam -o output_sorted.bam output.bam
samtools index output_sorted.bam output_sorted.bai

rm R1.fp.fastq
rm R2.fp.fastq
rm reference_genome.fasta
rm output.sam

##### copy results back to where they need to be #########
cp -aR * output_folder
```

You can view the output files created in the programs Tablet (Milne et al., 2012) (Figure 10) or Unipro UGENE (Okonechnikov et al., 2012) (

Figure 11).

Figure 10 A screenshot of the output (sorted_bam) file viewed in Tablet.

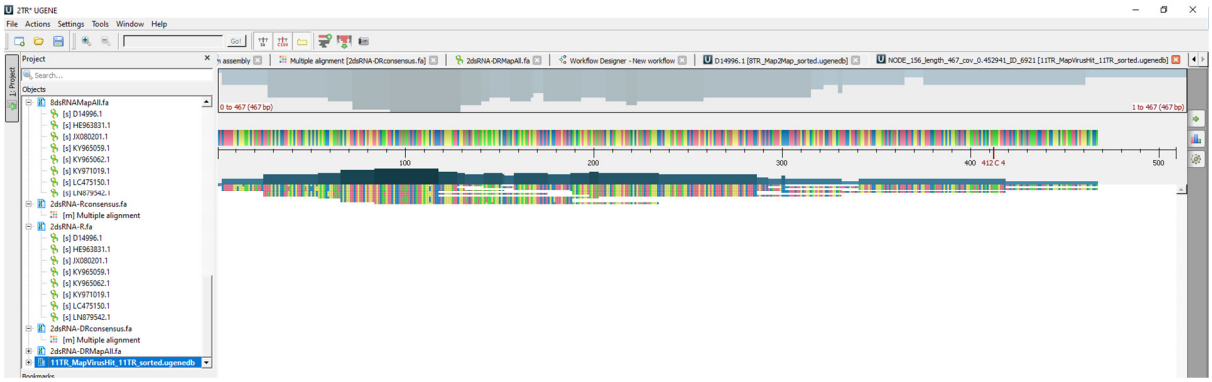

Figure 11 A screenshot of the output (sorted_bam) file viewed in UGENE

3.5 Depth and Coverage

It is important to check the depth and coverage of the reads mapped to ensure the genome has adequate coverage. I.e. A 10,000 bp viral genome with only 3 reads mapped to it, covering a 300bp region is NOT enough to be called a positive identification. A 10,000 bp viral genome with 80% genome coverage with 400 times depth, however, confers much more confidence. An example of a genome with good coverage is shown below (Figure 12).

Good coverage over the whole viral genome

Figure 12 A genome with good coverage over the entire genome shown in UGENE

You can check depth and coverage of bam files by:

- Load the module: module load SAMtools

- samtools coverage ouput.bam (name of bam file)

- The output will look like this:

```
#rname  startpos  endpos  numreads  covbases  coverage  meandepth
CP045952.1  1  5022086  1609496  4520045  90.0033  77.6183  37.1  40.4
```

3.6 Phylogeny analysis

Once the contigs have been confirmed and identified. A phylogenetic tree can now be created with your favourite programs. A maximum likelihood tree is usually recommended for the phylogeny analysis. Remember to consult the ICTV species demarcation criteria respective species.