microorganisms

MDPI

Review

A primer on the analysis of high-throughput sequencing data for detection of plant viruses

Denis Kutnjak[1,*,†], Lucie Tamisier[2,†], Ian Adams[3], Neil Boonham[4], Thierry Candresse[5], Michela Chiumenti[6], Kris De Jonghe[7], Jan Kreuze[8], Marie Lefebvre[5], Gonçalo Silva[9], Martha Malapi-Wight[10], Paolo Margaria[11], Irena Mavrič Pleško[12], Sam McGreig[3], Laura Miozzi[13], Benoit Remenant[14], Jean-Sebastien Reynard[15], Johan Rollin[2,16], Mike Rott[17], Olivier Schumpp[15], Sebastien Massart[2,†], Annelies Haegeman[7,†]

[1] Department of Biotechnology and Systems Biology, National Institute of Biology, Večna pot 111, 1000 Ljubljana, Slovenia
[2] Université de Liège, Gembloux Agro-Bio Tech, TERRA, Plant Pathology Laboratory, Passage des Déportés, 2, 5030 Gembloux, Belgium
[3] Fera Science limited, York YO411LZ, United Kingdom
[4] IAFRI, Newcastle University, King's Rd, Newcastle Upon Tyne NE1 7RU, UK
[5] Univ. Bordeaux, INRAE, UMR BFP, 33140, Villenave d'Ornon, France
[6] Institute for Sustainable Plant Protection, National Research Council, Via Amendola, 122/D, 70126 Bari, Italy
[7] Flanders Research Institute for Agriculture, Fisheries and Food, Plant Sciences Unit, Burg. Van Gansberghelaan 96, 9820 Merelbeke, Belgium
[8] International Potato Center (CIP), Avenida la Molina 1895, La Molina, Lima, Peru
[9] Natural Resources Institute, University of Greenwich, Central Avenue, Chatham Maritime, Kent ME4 4TB, UK
[10] USDA-APHIS-BRS, Biotechnology Risk Analysis Program, Riverdale, Maryland, USA
[11] Leibniz Institute-DSMZ, Inhoffenstrasse 7b, 38124 Braunschweig, Germany
[12] Agricultural Institute of Slovenia, Hacquetova ulica 17, 1000 Ljubljana, Slovenia
[13] Institute for Sustainable Plant Protection, National Research Council of Italy (IPSP-CNR), Torino, Strada delle Cacce 73, 10135 Torino, Italy
[14] ANSES Plant Health Laboratory, 7 rue Jean Dixméras, 49044 Angers Cedex 01, France
[15] Agroscope, Route de Duillier 50, 1260 Nyon, Switzerland
[16] DNAVision, Gosselies, Belgium
[17] ######
* Correspondence: denis.kutnjak@nib.si
† These authors contributed equality to this review.

Citation: Lastname, F.; Lastname, F.; Lastname, F. Title. *Microorganisms* **2021**, *9*, x. https://doi.org/10.3390/xxxxx

Academic Editor: Firstname Lastname

Received: date
Accepted: date
Published: date

Abstract: High-throughput sequencing (HTS) technologies have become indispensable tools assisting plant virus diagnostics and research thanks to their ability to detect any plant virus in a sample, without a prior knowledge. As HTS technologies are heavily relying on bioinformatics analysis of the huge amount of generated sequences, it is of utmost importance that researchers can rely on efficient and reliable bioinformatic tools and can understand the principles, advantages and disadvantages of the tools used. Here, we present a critical overview of the steps involved in HTS as employed for plant virus detection and virome characterization. We start from sample preparation and nucleic acid extraction as appropriate to the chosen HTS strategy, followed by basic data analysis requirements, an extensive overview of the in-depth data processing options and taxonomic classification of viral sequences detected. By presenting the bioinformatic tools and a detailed overview of the consecutive steps that can be used to implement a well-structured HTS data analysis in an easy and accessible way, this paper is targeted at both beginners and expert scientists engaging in HTS plant virome projects.

Keywords: plant virus; high-throughput sequencing; bioinformatics; detection; discovery;

1. Introduction

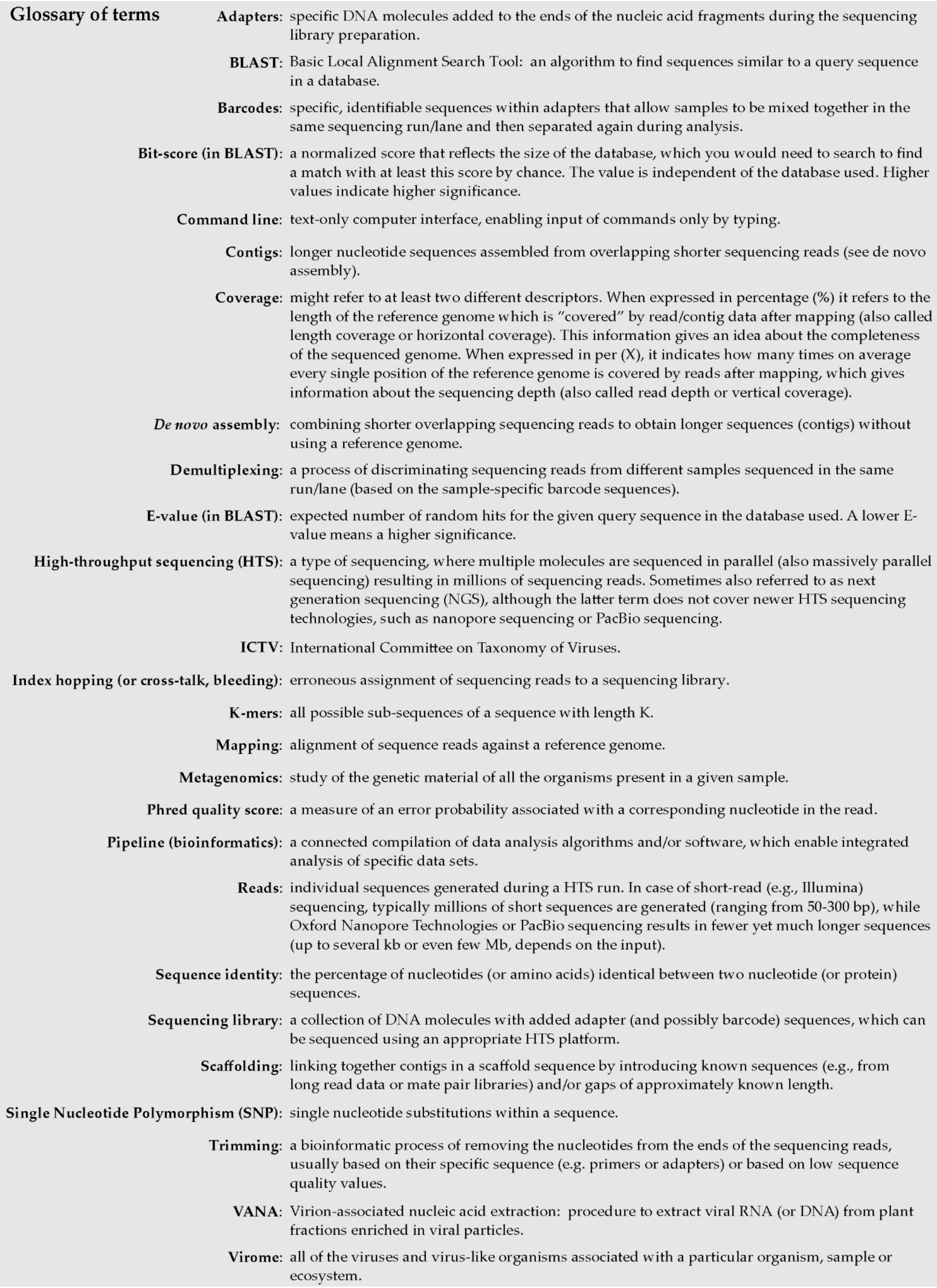
High-throughput sequencing (HTS) technologies have become an integral part of research and diagnostics toolbox in life sciences, including phytopathology and plant virology [1]. HTS enables the untargeted acquisition of extremely large amounts of sequence data from diverse sample types and thus represents an ideal and unique solution for generic detection of highly diverse viruses. In the past decade, sequencing prices have significantly decreased and the technology has become accessible to many more research and diagnostic labs. From the first uses of HTS for detection of plant viruses in 2009 [2–5], the use of this technology for detection of known and new plant viruses and the characterization of viromes in different plant species has escalated dramatically. Many different bioinformatics tools have been developed and different pipelines have been used to detect and identify plant viruses represented in HTS datasets. The variation in results associated with the use of different pipelines in different labs has highlighted the significance of understanding different approaches [6]. Arguably, one of the main challenges for less experienced users of HTS is to understand, select and properly use tools for the analysis of HTS data intended for detection and identification of plant virus sequences. In this review we aim to present the different and often complementary approaches used for analysis of HTS data for the detection of plant viruses. Here, we provide a short introduction to the laboratory work required and then describe the possible steps in data processing for detection of plant viruses, including: quality control and trimming of the sequences, *de novo* assembly, sequence similarity searches and taxonomic classification of the identified viral sequences. By including a short glossary (Figure 1), checklists and comparison tables, we aim to present the topic to the widest possible audience and thus encourage the use of HTS technologies by researchers with limited experience in the field.

2. What should I anticipate and how should I prepare?

Modern sequencing platforms can generate massive amounts of data, and not all laboratories wishing to use HTS in their projects have the necessary infrastructure and bioinformatics expertise, which, for example, is one of the main challenges identified for the adoption of these technologies in diagnostic laboratories [7]. The cost of the bioinformatics analysis in a HTS project was estimated to be around 15% of the total cost of a program (an example for whole genome analysis in cancer research), and includes the salary of the bioinformatician and cost of data storage [8].

Some commercial sequence analysis software is able to handle HTS data (see section 4.3.8), with dedicated modules for common operations (e.g., mapping and assembly). These software solutions are usually easy to use, regardless of the user's bioinformatics skill, but they are also quite expensive and might be limited for some analyses (specific applications). Furthermore, some "all in 1" viral-detection focused pipelines are available (see section 4.3.8), which require only limited bioinformatics knowledge or only the help of a skilled (bio)informatician at the installation stage.

However, to analyze HTS data, particularly for some specific applications, the use of dedicated bioinformatics software, without easy-to-use graphical user interface, is often needed to optimize time and efforts. These programs have in a large part been developed and optimized for the Linux platform, can be used in the command line only and so require specific computing skills. Considering the number of steps with the average HTS analysis pipeline and the number of samples studied, automation quickly becomes a priority. This can be achieved by writing scripts, as well as grouping and ordering all the steps of the analysis, which also require expertise in programing languages (e.g., shell, Python, R). Finally, for the interpretation of the analysis results, skills beyond pure bioinformatics are needed. A close collaboration between a bioinformatician and a plant virologist (or a plant virologist trained in bioinformatics) is needed to achieve a meaningful interpretation of the results.



Glossary of terms

Adapters: specific DNA molecules added to the ends of the nucleic acid fragments during the sequencing library preparation.

BLAST: Basic Local Alignment Search Tool: an algorithm to find sequences similar to a query sequence in a database.

Barcodes: specific, identifiable sequences within adapters that allow samples to be mixed together in the same sequencing run/lane and then separated again during analysis.

Bit-score (in BLAST): a normalized score that reflects the size of the database, which you would need to search to find a match with at least this score by chance. The value is independent of the database used. Higher values indicate higher significance.

Command line: text-only computer interface, enabling input of commands only by typing.

Contigs: longer nucleotide sequences assembled from overlapping shorter sequencing reads (see de novo assembly).

Coverage: might refer to at least two different descriptors. When expressed in percentage (%) it refers to the length of the reference genome which is "covered" by read/contig data after mapping (also called length coverage or horizontal coverage). This information gives an idea about the completeness of the sequenced genome. When expressed in per (X), it indicates how many times on average every single position of the reference genome is covered by reads after mapping, which gives information about the sequencing depth (also called read depth or vertical coverage).

***De novo* assembly:** combining shorter overlapping sequencing reads to obtain longer sequences (contigs) without using a reference genome.

Demultiplexing: a process of discriminating sequencing reads from different samples sequenced in the same run/lane (based on the sample-specific barcode sequences).

E-value (in BLAST): expected number of random hits for the given query sequence in the database used. A lower E-value means a higher significance.

High-throughput sequencing (HTS): a type of sequencing, where multiple molecules are sequenced in parallel (also massively parallel sequencing) resulting in millions of sequencing reads. Sometimes also referred to as next generation sequencing (NGS), although the latter term does not cover newer HTS sequencing technologies, such as nanopore sequencing or PacBio sequencing.

ICTV: International Committee on Taxonomy of Viruses.

Index hopping (or cross-talk, bleeding): erroneous assignment of sequencing reads to a sequencing library.

K-mers: all possible sub-sequences of a sequence with length K.

Mapping: alignment of sequence reads against a reference genome.

Metagenomics: study of the genetic material of all the organisms present in a given sample.

Phred quality score: a measure of an error probability associated with a corresponding nucleotide in the read.

Pipeline (bioinformatics): a connected compilation of data analysis algorithms and/or software, which enable integrated analysis of specific data sets.

Reads: individual sequences generated during a HTS run. In case of short-read (e.g., Illumina) sequencing, typically millions of short sequences are generated (ranging from 50-300 bp), while Oxford Nanopore Technologies or PacBio sequencing results in fewer yet much longer sequences (up to several kb or even few Mb, depends on the input).

Sequence identity: the percentage of nucleotides (or amino acids) identical between two nucleotide (or protein) sequences.

Sequencing library: a collection of DNA molecules with added adapter (and possibly barcode) sequences, which can be sequenced using an appropriate HTS platform.

Scaffolding: linking together contigs in a scaffold sequence by introducing known sequences (e.g., from long read data or mate pair libraries) and/or gaps of approximately known length.

Single Nucleotide Polymorphism (SNP): single nucleotide substitutions within a sequence.

Trimming: a bioinformatic process of removing the nucleotides from the ends of the sequencing reads, usually based on their specific sequence (e.g. primers or adapters) or based on low sequence quality values.

VANA: Virion-associated nucleic acid extraction: procedure to extract viral RNA (or DNA) from plant fractions enriched in viral particles.

Virome: all of the viruses and virus-like organisms associated with a particular organism, sample or ecosystem.

Figure 1. Glossary of terms commonly used in bioinformatics analysis of HTS data for plant virus detection.

Beyond the skills of users, IT resources must also be addressed. The amount of data generated by each project must be anticipated in order to have raw data storage space available beforehand, and to ensure that data is safely stored at least for several years after the end of projects. Depending on the sequencing platform, the total size of the raw data can become very large. For example, the Illumina NextSeq platform can generate from 120 to 300 Gbases (Gb) per run leading to file sizes varying between 39 and 170 GB depending on the read length. A stable and fast internet connection is often needed to facilitate efficient transfer of large data files. The computing resources also need to be anticipated. For time efficient analysis, it is often necessary to have a more powerful machine than an average workstation to run the various parts of pipelines, regardless of the software used. An alternative to the acquisition of a powerful computer is making use of online bioinformatics platforms and cloud computing solutions. These platforms generally have a structure adapted to the use of software making high demands on system resources (e.g., computing clusters). Many research centers or Universities host a Galaxy instance, which represent a very good alternative to the Linux platforms, in a more "user friendly" presentation.

3. Starting the project: How do I prepare samples and sequence nucleic acids?

Sampling, nucleic acids extraction, viral enrichment and sequencing library preparation are essential steps before HTS itself. Since these steps can influence the sequencing results, we briefly summarize here the most important considerations for some of these processes. Extensive description of how to control all of these steps is in preparation in forthcoming international guidelines for the use of HTS tests for the diagnostic of plant pests [9]. After obtaining the nucleic acids suitable for further analysis using HTS, approximate amount of sequence data required per each sample should be estimated according to the goals of the study. If external sequencing provider will perform HTS, this number, together with some general characteristic of the samples, should be communicated with the provider.

3.1. Input material and nucleic acids preparation

The extraction step separates the nucleic acids (including viral nucleic acids) from other cellular components. There are many methods that can be used to obtain high quality nucleic acids intended for HTS [10–13]. The efficiency of an extraction method is evaluated by the quantity of nucleic acids obtained, their integrity and the absence of contaminants that inhibit the enzymatic activities involved in the preparation of sequencing libraries. Irrespective of the chosen nucleic acid extraction procedure and library preparation methodology, it is recommended to collect several samples per plant to overcome the uneven distribution of viruses within the plant, especially in the case of low titer viruses. Different types of nucleic acids can be used as inputs for HTS, combined with different viral enrichment methods. No method is universal [11,14], each favors certain viral families or certain experimental objectives [15]. For example, total RNA or small RNA sequencing might be most straight-forward and universal to use for single samples. On the other hand, for sequencing of pools of many samples, or to optimize the detection of viruses with a low titer, methods that allow the enrichment of viral nucleic acids, such as Virion-Associated Nucleic Acids extraction (VANA) or the purification of double-stranded RNA might be preferred. The choice for one of the approaches should be based on the research question and study design. The purpose of the following paragraphs is to help making the most appropriate choices for sample preparation.

3.1.1. Total RNA/DNA

Extraction of total RNA and/or, to a lesser extent, DNA is a widely used approach for HTS analysis of plant tissues infected with viruses. Simple and robust, the method can be carried out according to several standard extraction protocols in solid phase or in liquid phase (Tan and Yiap, 2009) or using commercial kits (mostly based on silica-membrane or magnetic bead purification). Extraction and sequencing of total DNA can be sometimes used specifically for the detection of DNA viruses, while sequencing of total RNA is a very generic approach and can be used for detection of all types of DNA and RNA viruses and viroids [15]. The high abundance of nucleic acids from the host plant co-extracted with viral nucleic acids can greatly limit the sequencing sensitivity. The relative proportion of viral sequences in the total extracted RNA can be increased by the depletion of the plant ribosomal RNA [16,17] and the proportion of sequences of circular DNA viruses in extracted DNA can be enriched by rolling circle amplification [18–20].

3.1.2 Small RNA (sRNA)

The plant immune system responds to the presence of viruses by activating a defense response which leads to the cleavage of double stranded forms of viral RNA into small RNAs (sRNA) of 21 and 22 nucleotides (nt) as well as, more marginally, of 24 nt [21]. The analysis of sRNAs facilitates the reconstruction of the complete genomes of infecting RNA and DNA viruses or viroids, as well as those of integrated endogenous viral elements (EVEs) if they are transcribed [2,15,22,23]. Since sRNAs are more stable than longer RNA molecules, the method is promising for use in old or even ancient plant samples [24] and since only very short reads are needed to sequence sRNAs, the method is relatively cost efficient. On the other hand, d*e novo* assembly from short sequences is complex and might lead to chimeric sequences in case of multiple infections with different virus strains [25] and for the same reason, pooled samples used in metagenomic studies including large number of plants are not recommended to be analyzed with sRNA sequencing. Due to their short lengths, analyses of recombination events on a read level are also not feasible with sRNA [22].

3.1.3 Virion-associated nucleic acids (VANA)

The extraction of Virion-Associated Nucleic Acids (VANA) enriches the samples in nucleic acids of viral origin by semi-purifying the viral particles by ultracentrifugation. Viral particles are separated from most of the organelles and plant debris by one or two differential ultracentrifugation cycles depending on the viral family and the plant material. After purification of the particles, and a nuclease treatment to degrade non protected nucleic acids, the viral nucleic acids are extracted according to a standard extraction protocol also used for the extraction of total RNA/DNA. Initially developed for the biochemical characterization of viral particles in the 1970s, VANA was used in pioneering studies of prospecting for viral diversity in wild asymptomatic plants before the advent of HTS [26,27]. The approach was then extended to the preparation of nucleic acids intended for HTS [28,29]. It achieves balanced enrichment in high quality viral RNA and DNA and allows the use of up to several hundred grams of starting material. However, it is based on the stability of the viral particles mainly determined by the pH and the concentration of salts in the extraction buffer. Unsuitable for high throughput, and relying on numerous laboratory operations, the approach only identifies the encapsidated viral nucleic acids as well as the viruses of the *Endornaviridae* family, devoid of capsids but encapsulated in membranous vesicles [28,30]. Moreover, certain viral families are difficult to purify and VANA is also not the method of choice for extraction of viruses from plants with high content of phenolic and polysaccharide compounds [31].

3.1.4 Double-stranded RNA

The majority of plant viruses have RNA genomes, accounting for 75% of the total number of viruses reported [32]. While plants do not produce large quantities of double-stranded (ds)RNAs, RNA viruses generate high molecular weight dsRNA structures during replication, so their enrichment is a popular strategy used for virus diagnostics

[10,13,33,34]. Extraction of dsRNA purifies nucleic acids from double-stranded RNA viruses, but also from most single-stranded RNA viruses, viroids as well as from some DNA viruses [35–38]. This approach allows the detection of a very wide range of RNA virus species [30,39]. Sequencing of dsRNA is likely not the most effective method for detection of negative sense single stranded RNA viruses [37]. It is also a laborious approach, even if, a number of modified protocols have been developed to overcome this limitation [13,34,40–42].

3.2 Library preparation and sequencing

Following nucleic acid extraction, different methods have been developed for library preparation using commercially available kits and automated systems. As inputs, the extracted and possibly virus-enriched nucleic acids described in previous paragraphs can be used. The type of the library preparation and exact protocol is dependent on the input nucleic acids (e.g., total RNA or DNA, sRNA, dsRNA). Specific libraries are prepared for different HTS platforms. The library preparation step usually consists of shearing the nucleic acids to the size appropriate for intended sequencing platform and the ligation of short oligonucleotides (adaptors) at one or both extremities of the nucleic acids in order to allow the sequencing. There are two main groups of HTS platforms: (i) short read HTS (also termed next generation sequencing – NGS), producing reads up to several hundred nucleotides, and (ii) long read HTS (also termed single molecule sequencing – SMS), producing reads up to hundreds of kilobases (kb). Currently, the most commonly used sequencing platform is Illumina (short read HTS), and, for long read HTS, Pacific Biosciences (PacBio) and Oxford Nanopore Technologies. Nanopore sequencing is rapidly developing and is expected to be more widely used in the future [43]. Most of the available protocols recommend assessing the quality and quantity of the nucleic acids before library preparation. The integrity and purity of the nucleic acids can be assessed using spectrophotometric and fluorescence-based assays. For some enrichment approaches (e.g., VANA, dsRNA extraction), the concentrations of the obtained nucleic acids frequently are below the input required for library preparation so that a random amplification step is required prior to library construction [13].

Several samples can be pooled and sequenced in the same sequencing run (multiplexing). In this case, the oligonucleotides ligated to the nucleic acids during library preparation also include specific sequences corresponding to barcodes unique for each sample. After sequencing, the reads are allocated to the appropriate samples according to the barcodes used. Most commonly, sequencing reads are contained in a fastq file format, which also contains some technology-specific descriptors and nucleotide quality values. The fastq files represent an input for the bioinformatics analysis described in the following paragraphs.

4. How do I analyze the data?

The first step towards successfully annotating viral reads or contigs is to perform a quality control of the raw HTS data. After quality control of raw data and prior to any downstream analysis, it is important to perform pre-processing steps including trimming low-quality bases, removing adapter sequences and discarding very short and low-quality reads. These steps are explained in more detail in section 4.1. Afterwards, the quality of pre-processed reads can be checked again. Resulting reads can then be analyzed for similarity with known viral sequences in several ways as depicted in Figure 2. Trimmed and filtered reads can be analyzed directly, or they can be first assembled into longer contigs (section 4.2). The contigs can then be analyzed for similarity with viral sequences. Specifically, reads or contigs resulting from assembly can be used directly for similarity searches against sequence or domain databases, or can be first mapped to a host reference genome, if available, so that sequences originating from the host can be removed. Reads that do not map to the host genome can optionally be assembled into longer contigs. The contigs obtained from the *de novo* assembly step can then be annotated using different strategies depending if they represent known or putative novel viruses. Contigs associated with known viruses can be mapped to the corresponding viral reference genome. In the case of putative novel viruses, a search of conserved motifs in the translated theoretical protein sequences can be performed alongside similarity searches at the protein level using BLASTx or DIAMOND tools as explained in section 4.3. Results of those analyses need to be carefully inspected and further analyses often need to be performed for correct taxonomic classification of the sequences (section 4.3.7). The described steps can be performed using the tools indicated in the flow chart (Figure 2) or other available tools. Finally, the same analyses can also be performed using user-friendly free software with graphical user interfaces (GUI) available online or using commercial software as described in section 4.3.8.

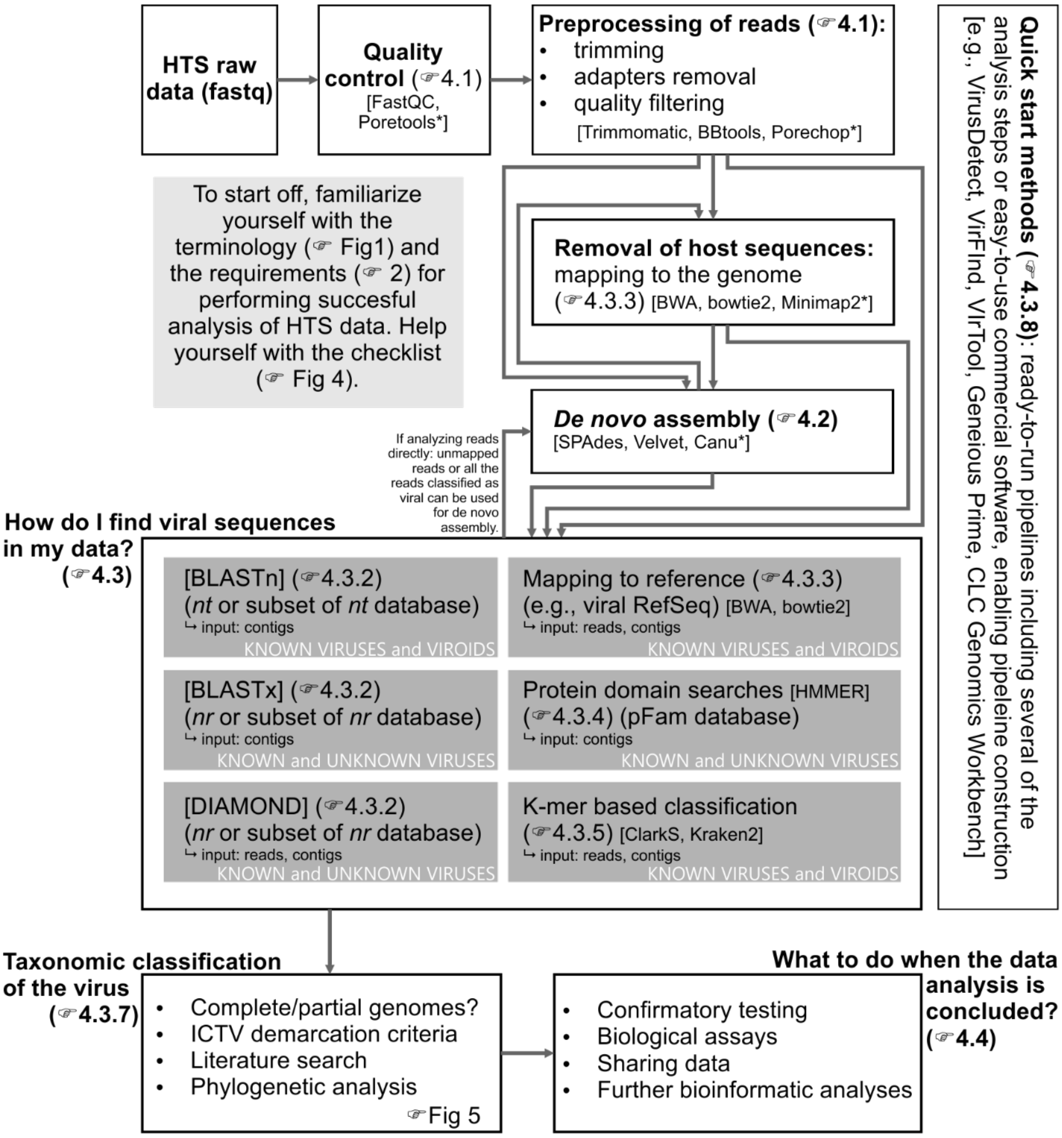

HTS raw data (fastq)
Quality control (☞4.1) [FastQC, Poretools*]
Preprocessing of reads (☞4.1):
• trimming
• adapters removal
• quality filtering
[Trimmomatic, BBtools, Porechop*]
Quick start methods (☞4.3.8): ready-to-run pipelines including several of the analysis steps or easy-to-use commercial software, enabling pipeleine construction [e.g., VirusDetect, VirFind, VirTool, Geneious Prime, CLC Genomics Workbench]
To start off, familiarize yourself with the terminology (☞ Fig1) and the requirements (☞ 2) for performing succesful analysis of HTS data. Help yourself with the checklist (☞ Fig 4).
Removal of host sequences: mapping to the genome (☞4.3.3) [BWA, bowtie2, Minimap2*]
De novo assembly (☞4.2) [SPAdes, Velvet, Canu*]
If analyzing reads directly: unmapped reads or all the reads classified as viral can be used for de novo assembly.
How do I find viral sequences in my data? (☞4.3)
[BLASTn] (☞4.3.2) (nt or subset of nt database) ↳ input: contigs
KNOWN VIRUSES and VIROIDS
Mapping to reference (☞4.3.3) (e.g., viral RefSeq) [BWA, bowtie2] ↳ input: reads, contigs
KNOWN VIRUSES and VIROIDS
[BLASTx] (☞4.3.2) (nr or subset of nr database) ↳ input: contigs
KNOWN and UNKNOWN VIRUSES
Protein domain searches [HMMER] (☞4.3.4) (pFam database) ↳ input: contigs
KNOWN and UNKNOWN VIRUSES
[DIAMOND] (☞4.3.2) (nr or subset of nr database) ↳ input: reads, contigs
KNOWN and UNKNOWN VIRUSES
K-mer based classification (☞4.3.5) [ClarkS, Kraken2] ↳ input: reads, contigs
KNOWN VIRUSES and VIROIDS
Taxonomic classification of the virus (☞4.3.7)
• Complete/partial genomes?
• ICTV demarcation criteria
• Literature search
• Phylogenetic analysis
☞Fig 5
What to do when the data analysis is concluded? (☞4.4)
• Confirmatory testing
• Biological assays
• Sharing data
• Further bioinformatic analyses

Figure 2: Flowchart representing different approaches for the analysis of HTS data for the detection of plant viruses. Boxes represent different steps in data analysis and interpretation. Arrows connect different possible sequences of the analysis steps. As an example, a non-exhaustive list of possible analysis tools is added in the square brackets at each of the analysis steps. Tools designated with * are intended for use with long-read or, specifically, nanopore sequencing data. Pointing hands lead to the text sections (or figures) with more detailed description of the corresponding steps.

4.1. Demultiplexing, quality control and trimming

Each sequencing platform produces a series of quality metrics associated with the data produced from each sequencing run. A discussion of the metrics with the sequencing data provider is important before accepting any sequencing data.

If the run was successful, the first step is the demultiplexing of barcoded samples, which usually carried out using the sequencing platform software or performed by the sequencing data provider. In the event that data has not been demultiplexed, third party tools such as Cutadapt [44] can be used to demultiplex the Illumina data by looking for specific barcode sequences present in the samples. Alternatively, demultiplexing tools developed by the sequencing platform provider are frequently accessible as stand-alone tools, such as Illumina's bcl2fastq software [45], or Oxford Nanopore Technologies' guppy barcoder script [46].

Adapter sequences introduced during the library preparation process need to be removed. Tools such as Cutadapt [44], Trimmomatic [47] and Porechop [48] or NanoFilt [49] can be used to carry out this process, with the latter two working specifically for data generated using nanopore sequencers. At this step, contaminant filtering for synthetic molecules and/or spike-in is also recommended.

Sequencing data is usually provided in the fastq format, which consists of four lines per sequence [50], including sequence identifier, raw nucleotide sequence, a separator line (containing + sign) and sequence quality values. Quality values present in a fastq file represent Phred quality scores, which are encoded as ASCII characters. The quality score (Q) associated with each nucleotide represents the estimated probability of an error. For example, a quality score of 0 represents a 100% chance of an error, Q10 = 10% chance of an error, Q20 = 1% chance of an error etc.

Nucleotides with a low-quality score should be removed to ensure that only high accuracy bases remain. With Illumina data, values such as Q20 (1% error) and Q30 (0.1% error) are often used when trimming data, but this value depends on the application and the sequencing platform used. If accuracy is of the utmost importance (e.g., for detection of SNPs), selecting a higher quality score will be beneficial. If accuracy is less important (e.g., for detection of virus), then relaxing constraints on quality when trimming will allow more data to be available for downstream applications.

Quality control reports can be generated by tools such as FastQC [51], MultiQC [52], or, specifically for nanopore sequencing data, Poretools [53] or NanoStat [49]. This allows for the visual inspection of metrics such as per base sequence quality, sequence length distribution and GC (guanine-cytosine) content. These reports can be generated both before and after trimming, to assess the impact of trimming on different quality parameters. A number of tools exist to trim sequencing reads based on quality scores, sequence length or other metrics. These include, but are not limited to, Sickle [54], Trimmomatic [47], Cutadapt [44], BBDuk (https://sourceforge.net/projects/bbmap/) and NanoFilt for nanopore sequencing data [49]. Illumina data, particularly longer MiSeq reads, suffer from lower quality towards the 3′ end of the read. Many trimming strategies start at the 3′ end of such reads and determine the position at which the quality (or the average quality in a region) is high enough to keep.

The order in which these processes are carried out can vary, and some tools can be used to carry out multiple steps at the same time. The final output should be a series of demultiplexed samples with reads that have an acceptable sequence quality and no longer contain sequences added during the sequencing process (e.g., adapters, barcodes).

4.2 De novo assembly

Sequencing technologies have improved in quality and amount of the data generated in the last ten years. However, up to now, it is not possible to obtain the full-length genome of many organisms in a single high quality read. HTS technologies provide us with shorter (e.g. Illumina) or longer (e.g. Oxford Nanopore Technologies, PacBio) sequence fragments, which usually need to be assembled *in silico* to reconstruct complete or near-complete genomes. Compared to bacteria or eukaryotes, viral genomes are simpler and smaller. Nevertheless, high mutation rates and consequently the great diversity of some viral populations [55] might represent a challenge for *in silico* genome reconstruction. Assembling a genome is like solving a "Jigsaw puzzle". Similar to a puzzle there could be pieces fitting together (overlapping reads), missing pieces (regions with low coverage, sequencing bias) and damaged parts (sequencing errors). The process for which individual reads sharing sequence similarity are merged to form longer fragments is named *de novo* sequence assembly and the nucleotide fragments obtained through this process are called contigs or contiguous sequences [56].

Depending on the platform used, the sequence reads can be large or small and from the computational point of view, different intrinsic features of these two types of output, led to the development of two major groups of assembly algorithms: (i) de Bruijn graph (DBG) and (ii) the overlap-layout-consensus (OLC) methods. In the first case, DBGs are constructed using k-mers, which are substring of the reads of length k; whereas for OLC, the overlap graphs are constructed directly from reads, eliminating the redundant ones. The use of k-mers is more widely applied for the assembly of short reads, whilst the OLC approach is most appropriate for long read data [56,57].

For short HTS reads, many de Bruijn graph assemblers are available, such as SOAPdenovo2 [58], ALLPATHS-LG [59], ABySS [60], Velvet [61], IDBA-UI [62] and (rna)SPAdes [63–65]. One of the first and most widely used and cited assembler [66] in viral metagenomics [67], is the open-source software Velvet, followed by the more user-friendly and commercially-available CLC Genomics Workbench (https://digitalinsights.qiagen.com) and Geneious Prime (https://www.geneious.com). The latter has the advantage of providing a graphical interface for command-line assembly programs like Velvet and Spades.

Different factors can positively influence the quality of the *de novo* assembly, e.g., a preliminary filtering step to eliminate the genomic host plant reads [23] or the selection of appropriate k-mer values based on the read length [67]. Moreover, approaches in which *de novo* assemblies using different k-mer values are generated and then reassembled can generally improve the completeness of *de novo* genome assemblies, but this can be a laborious and computationally lengthy process. Usually higher sequencing depth and a higher fraction of viral reads in the dataset will positively affect the completeness of assembled viral genomes, however, extremely high coverage might have a negative effect on the completeness of the assembly when using some assemblers, thus, in such cases, assembly of subsampled data might give better results [15]. Since reads of some viruses can be present in a very low number, it is important not to set too low cut-offs for contigs length [67], e.g., a number around or slightly above the 2x length of an average read length is recommended. Finally, the use of an additional scaffolding step when using paired-end data can sometimes further increase the length of the contigs. Nevertheless, despite improvements in *de novo* assembly algorithms, 3′ and 5′ ends of viral genomes usually cannot be obtained in full through *de novo* assembly.

Although long read HTS platforms can produce reads close to full-length viral genomes, a major issue that could affect the *de novo* assembly step is the higher error rate (5-15%) of these technologies [68]. Long-read assemblers can algorithmically correct base errors before/when building contigs. PBcR [69], Canu [70], Falcon [71] and Pomoxis [72] are some of the OLC-based *de novo* assemblers available. Long read nanopore sequencing has recently been successfully applied to virus discovery, detection and reconstruction of virus genomes, in these studies, Canu is the most cited assembler [73–76].

Contigs generated by *de novo* assembly can be used in subsequent similarity searches and finally viral contigs can be used for phylogenetic or recombination analysis. If this is so, it is important to check the quality of such contigs by mapping the trimmed reads (explained in section 4.3) to the viral contigs followed by visual inspection of the mapping and checking the completeness of expected open reading frames contained in this contigs.

For contigs generated by *de novo* assembly of nanopore sequencing reads additional quality checking steps might be needed such as assembly polishing [75] or correction of the consensus sequences using quality data of mapping reads [76].

When the presence of specific viruses is already known, viral genomes can be reconstructed by mapping the reads (explained in section 4.3) to the closest reference sequences obtained from sequence databases (after initial similarity searches, section 4.3). This is then followed by the extraction of new consensus sequence from the mapping, an approach known as reference guided assembly. Sometimes, parts of the viral genomes are obtained by *de novo* assembly and parts through reference guided assembly; such an approach is also known as combined assembly.

4.3 How do I find and classify viral sequences in my data?

Identification of viral reads/contigs in massive datasets produced after HTS is most frequently performed by comparing sequences against known and annotated sequences in databases. Because longer sequences in almost all cases improve the ability to identify similarities regardless of the method or databases used, an assembly of quality checked raw reads is generally recommended prior to similarity searches. At the same time, a prior assembly will also generally reduce the computing time needed for the similarity search steps as up to millions of reads can be assembled in a single contig. Annotation of HTS reads, or contigs, on the basis of similarity with known viral sequences can be performed using three main strategies: homology searches with tools such as BLAST [77], read/contig mapping against reference viral genomes using tools such as BWA [78] and the search for encoded, conserved protein motifs using tools based on Hidden Markov Models (HMMs) such as HMMER [79]. Each of these approaches and, in turn, each of the specific programs used to perform them, has advantages and drawbacks. In many cases, they should be seen as complementary rather than mutually exclusive possibilities. Several additional alternatives have also been proposed. For example, the use of e-probes (short unique pathogen-specific reference sequences) [80] or the analysis of the frequency of specific k-mer sequences (see section 4.3.5). A summary of tools commonly used for similarity searches is presented in Table 1.

4.3.1 Databases

The database(s) against which sequences are compared is/are of utmost importance for the efficiency and completeness of the annotation process. The more complete the collection of viral sequences, the greater the likelihood of detecting and identifying the presence of a virus. For BLAST and BLAST-like approaches, the most used databases are the non-redundant nucleotide database (nr/nt, named also just nt) hosted by NCBI, the non-redundant GenBank protein database (nr) or the viral RefSeq database. The GenBank non-redundant nucleotide and protein databases are the most comprehensive and most frequently updated public databases, limiting the time from discovery of a novel virus to its availability for comparisons (provided the local version of these databases is also regularly updated). However, the size of these databases has the drawback of increasing the computing time/power needed to perform a comparison. The reduced viral RefSeq database has the benefit of better annotation/curation at the expense of the number of included sequences and of less frequent updates. For read mapping approaches, smaller dedicated databases are generally used, such as a subset of all viral sequences from the NCBI nt database, viral RefSeq or a smaller, locally developed and curated database (for example, one or several isolates of every virus known to infect the crop of interest). For conserved protein motifs searches, the most common databases are PFAM [81] and CDD [82]. Identification of viral sequences is critically dependent upon the quality of the database(s) used. For example, some plant derived proteins might also be misidentified as viral if only a virus sequence database is used for similarity searches, because some viral proteins are related to plant encoded proteins. Typical examples are heat shock proteins (*i.e.*,. Hsp70) proteins found in closteroviruses [83], or reverse transcriptase proteins of *Caulimoviridae* that have homologs among retrotransposons. Wrongly annotated sequences in the public databases can also lead to erroneous annotations.

Table 1. Summary of the most commonly used similarity search strategies with advantages and limitations for each of the strategies.

Tool name	Advantages	Limits and considerations	Important thresholds
BLASTx or BLASTn	High accuracy	Slow, intensive use of computing power if large database is used, BLASTx needed for detection of divergent novel viruses, BLASTn needed for detection of viroids and noncoding regions of viral genomes or satellites; performance improved by prior assembly of contigs.	Minimum percentage of identity; length of identified region of similarity; minimal e-value, bit-score.
MegaBLAST	Faster than BLASTn, handles longer sequences	Less sensitive than BLASTn, only useful for detection of nucleotide sequences very similar to the ones in the used database; performance improved by prior assembly of contigs.	Minimum percentage of identity; length of identified region of similarity; minimal e-value, bit-score.
BLASTp	High accuracy	Slow, need to translate nucleotide sequences to proteins first; performance improved by prior assembly of contigs; not applicable for viroids or noncoding regions of viral genomes or satellites.	Minimum percentage of identity; length of identified region of similarity; minimal e-value, bit-score.
DIAMOND	Faster than BLASTx	Less sensitive, annotation less accurate than BLAST; performance improved by prior assembly of contigs; only available for searches against protein databases; not applicable for viroids or noncoding regions of viral genomes or satellites.	Minimum percentage of identity; length of identified region of similarity; minimal e-value, bit-score; use sensitive mode.
Burrows-Wheeler transform-based mapping algorithms (e.g., BWA or Bowtie2)	Does not require prior assembly of contigs, high sensitivity for short sequences	Only allows detection of known agents. Difficult to adjust mapping stringency to (1) allow detection of divergent isolates while (2) avoiding cross-mapping between related agents; prior assembly of contigs reduces cross-mapping between related agents.	Mapping stringency (e.g., mismatch penalties, gap open/extension penalties, percent of read length matching reference, minimum percentage of identity…)
HMMER or HMM-Scan	High efficiency for detection of distant homologs	Annotation more complex for protein families shared between cellular organisms and viruses; not applicable for viroids or noncoding regions of viral genomes or satellites.	Minimal e-value.
K-mer based classification algorithms (Kraken or Taxonomer)	Fast	Requires large computer memory; accuracy may be limited for the shorter genomes of plant viruses; the confidence scoring of the results is not straight forward.	C/Q ratio for Kraken (advise the manual).

4.3.2 BLAST and BLAST-like approaches

BLAST programs are the most widely used and among the most accurate in detecting sequence similarity [84]. The BLAST suite [85] comprises different algorithms, each with its own use:

1. BLASTn can be used to compare a nucleotide sequence with a nucleotide database. It is less computationally intensive than BLASTx, but because of the higher divergence rate of nucleotide sequences, it is less efficient for the annotation of novel viruses not represented in the database used.

2. BLASTp can be used to compare a protein sequence with a database of protein sequences.
3. BLASTx can be used to compare a nucleotide sequence translated in all six reading frames with a database of protein sequences. While computationally intensive, it is the most efficient BLAST program for the annotation of novel viruses.
4. tBLASTn can be used to compare a protein sequence with all six possible reading frames of a nucleotide database and is often used to identify proteins in new, unannotated genomes.
5. tBLASTx can be used to compare all six reading frames of a nucleotide sequence with all six reading frames of a nucleotide database. It is the costliest in computation time.
6. MegaBLAST can be used to compare nucleotide sequences expected to be already present or closely related to those in a nucleotide database. It can be much faster than BLASTn and is able to handle much longer sequences but deals less efficiently with very divergent sequences.

Short sequences may lead to false positives in BLAST searches and for this reason, other approaches should be preferred for very short reads or contigs. All BLAST programs return a table of results, which contain several parameters, among which some are particularly important to check: the identity threshold (threshold for the % of identical nucleotides between the query sequence and a hit in a database), e-value (expected number of random hits in the used database for a given query sequence) and query coverage (% of the query sequence covered by the database hit). It is very important to consider that some of these values depend on the size of the database used and that the use of too stringent parameters (e.g. identity threshold >85% and e-value smaller than 10^{-10}) may lead to a failure to detect some divergent viruses [6]. BLAST is very widely used, but remains, in the case of millions/billions of reads analyses, a time-consuming algorithm. Restricting the database used to specific taxa (e.g., viruses) can speed up BLAST searches but care should be taken as this frequently leads to the identification viral reads which on closer examination, using complete databases, are in fact host sequences (e.g., plant sequences). An extremely fast but considerably less sensitive alternative to BLAST is BLAT (BLAST Like Alignment Tool) [86]. Another faster alternative to BLASTx is DIAMOND [87] which runs at 500-20,000x the speed of BLAST, while maintaining a high level of sensitivity, especially if using the sensitive mode. The DIAMOND annotations have however been observed to be less optimal in virus species identification than BLAST ones (ML & TC personal observations).

4.3.3 Mapping reads (or contigs) to reference database

Mapping tools are commonly used as a filtering step to remove host genome sequences or as a complement to similarity searches on short nucleotide sequences. Reads originating from the host genome can be partially removed by mapping the complete dataset to reference genomic sequences of corresponding host (if available) and then using only unmapped reads for further analyses. A reference genome sequence of the host must be chosen carefully, since it can affect the analysis. Choosing divergent variety/genotype of the host might reduce the efficiency of the host reads removal. Furthermore, reference host genomes might contain contaminating or genome-integrated viral sequences, thus, some viral reads can be lost in this step.

Mapping tools can be also used to perform the alignment of reads or contigs against a reference viral database (e.g., NCBI Viral RefSeq database or a custom developed database containing one or more complete or partial genomes). In comparison to BLAST programs, most of the mapping tools such as Bowtie2 [88] or BWA [78] build an index for the reference genome or the reads, increasing the speed of the analysis if used against a limited, virus-specific database. The mapping strategy is potentially more sensitive to detect viruses with low number of reads in analyzed datasets [6], in particular when using 21-24 nt sRNA sequences. Consequently, it is also sensitive to cross-sample contamination due to index-hopping, which may require the development of strategies to set a positivity threshold. On the other hand, mapping strategies are inefficient at detecting novel viruses or viroids that are absent from the database used. Mapping stringency parameters (see Table 1) critically affect the outcome of the analyses and should be optimized keeping in mind the objective of the experiment. Too stringent parameters may result in the failure to detect divergent viral isolates. Too relaxed parameters may also give rise to erroneous results, through the mapping of related host genes on a viral genome or through cross-mapping of reads of a virus on the genome of a related virus. It is therefore highly recommended to carefully analyze mapping results. An efficient strategy, besides counting the number of mapped reads on a particular reference genome considers the portion of this genome covered by the mapped reads, the percentage of similarity between mapped reads and the reference or other similar indicators to eliminate potential false positive results. Including suitable reference samples as controls during sample preparation and sequencing can help to eliminate such errors [9]. Similar to reads, contigs generated by *de novo* assembly can also be mapped to the reference databases. Due to the greater length of the contigs, less erroneous mapping results are expected. However, the same recommendations for careful inspection of mapping results apply.

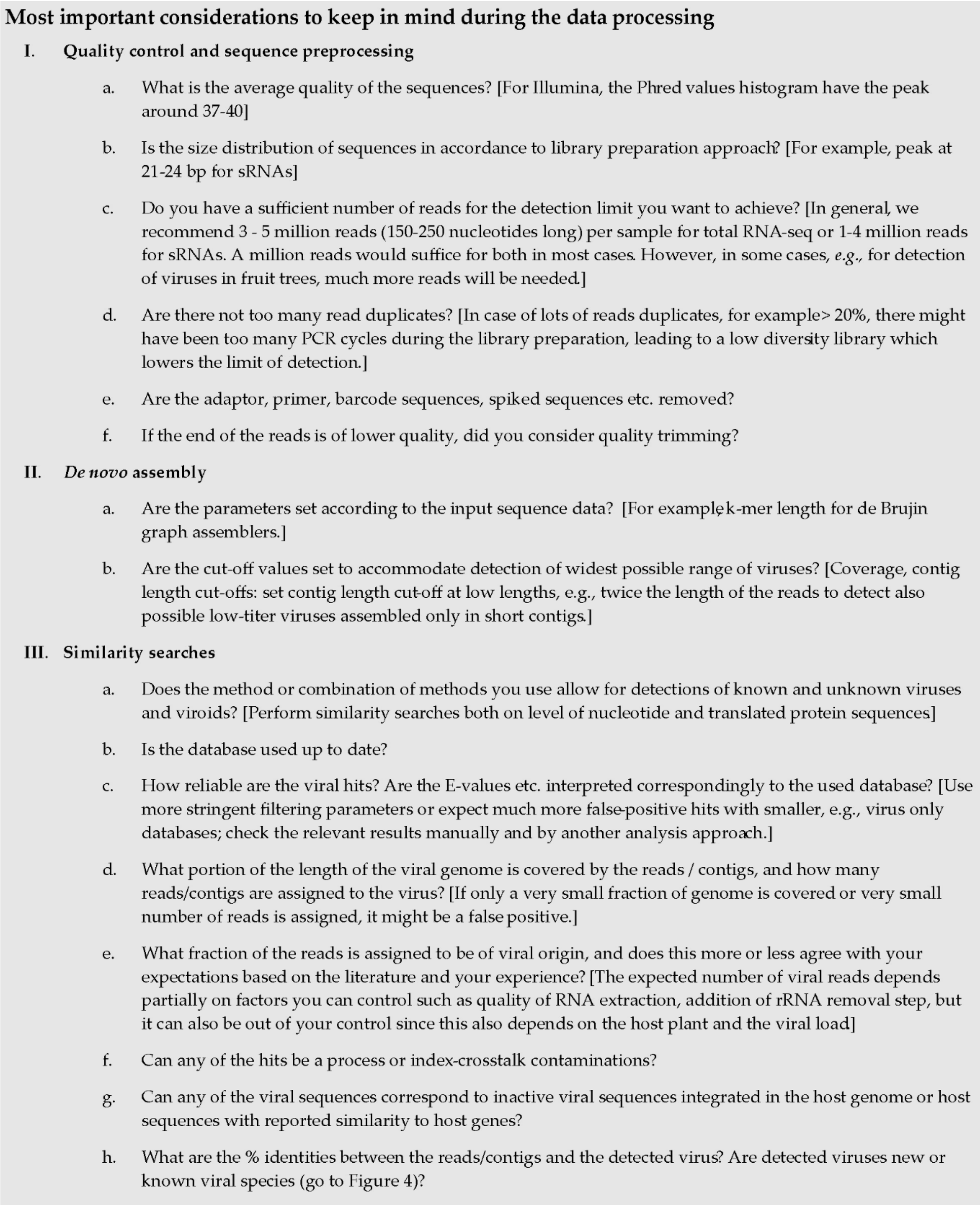
4.3.4 Protein domain searches

Searching for known viral domains by matching translated protein sequences of reads/contigs with Hidden Markov Models (HMMs) of known protein domains using programs such as HMMER [89] or HMMScan is a popular alternative to BLASTx. With this method, sequences are first translated in all possible reading frames and the translated protein sequences are compared to a database of conserved protein motifs such as PFAM [81] or CDD [82]. These approaches are faster than BLAST-based homology searches and more effective than mapping or BLAST searches for the detection of very distant homologs [90] and therefore, possibly for the detection of novel, very divergent viruses. Like with BLAST, a significance e-value is calculated, allowing the evaluation of the significance of a match. This e-value can be used to filter results, striking a balance between low values and the reporting of false-positives, and high values and the failure to detect a divergent virus.

4.3.5 K-mer approaches and machine learning-based approaches

Nucleotide k-mer-based approaches can be used to annotate sequences based on the presence and frequency of specific k-mers. Comparing these frequencies is computationally less demanding and faster than sequence alignment but requires a lot of computer memory. Even if most of the k-mer-based classification tools, such as Kraken [91,92], Kaiju [93] or Taxonomer [94], are not dedicated toward detection of plant viruses, they can be used for such purpose. Kodoja [95] uses a combination of such tools for the taxonomic classification of plant viruses in metagenomic data. Most of the tools are not very user friendly and the use of k-mer tools for plant virus detection is fairly new, thus some questions remain to be answered, e.g., the usability of k-mer tools on small RNA data sets [95].

Methods based on machine learning are being developed for detection of viral sequences in metagenomics datasets. Several tools have already been published, e.g., ViraMiner [96], DeepVirFinder [97] or Virnet [98] for human virus detection purpose. Given a metagenome with known composition, a machine learning approaches attempt to find some meaningful patterns that allow to differentiate the host from the virus. When unknown metagenome dataset is provided, the software should be able to discriminate virus sequences from host sequences using the learnt pattern. Machine learning tools are new in this field, thus, we still lack their in-depth comparison with the more known approaches discussed above.



Most important considerations to keep in mind during the data processing

I. Quality control and sequence preprocessing

a. What is the average quality of the sequences? [For Illumina, the Phred values histogram have the peak around 37-40]

b. Is the size distribution of sequences in accordance to library preparation approach? [For example, peak at 21-24 bp for sRNAs]

c. Do you have a sufficient number of reads for the detection limit you want to achieve? [In general, we recommend 3 - 5 million reads (150-250 nucleotides long) per sample for total RNA-seq or 1-4 million reads for sRNAs. A million reads would suffice for both in most cases. However, in some cases, *e.g.*, for detection of viruses in fruit trees, much more reads will be needed.]

d. Are there not too many read duplicates? [In case of lots of reads duplicates, for example> 20%, there might have been too many PCR cycles during the library preparation, leading to a low diversity library which lowers the limit of detection.]

e. Are the adaptor, primer, barcode sequences, spiked sequences etc. removed?

f. If the end of the reads is of lower quality, did you consider quality trimming?

II. *De novo* assembly

a. Are the parameters set according to the input sequence data? [For example, k-mer length for de Brujin graph assemblers.]

b. Are the cut-off values set to accommodate detection of widest possible range of viruses? [Coverage, contig length cut-offs: set contig length cut-off at low lengths, e.g., twice the length of the reads to detect also possible low-titer viruses assembled only in short contigs.]

III. Similarity searches

a. Does the method or combination of methods you use allow for detections of known and unknown viruses and viroids? [Perform similarity searches both on level of nucleotide and translated protein sequences]

b. Is the database used up to date?

c. How reliable are the viral hits? Are the E-values etc. interpreted correspondingly to the used database? [Use more stringent filtering parameters or expect much more false-positive hits with smaller, e.g., virus only databases; check the relevant results manually and by another analysis approach.]

d. What portion of the length of the viral genome is covered by the reads / contigs, and how many reads/contigs are assigned to the virus? [If only a very small fraction of genome is covered or very small number of reads is assigned, it might be a false positive.]

e. What fraction of the reads is assigned to be of viral origin, and does this more or less agree with your expectations based on the literature and your experience? [The expected number of viral reads depends partially on factors you can control such as quality of RNA extraction, addition of rRNA removal step, but it can also be out of your control since this also depends on the host plant and the viral load]

f. Can any of the hits be a process or index-crosstalk contaminations?

g. Can any of the viral sequences correspond to inactive viral sequences integrated in the host genome or host sequences with reported similarity to host genes?

h. What are the % identities between the reads/contigs and the detected virus? Are detected viruses new or known viral species (go to Figure 4)?

Figure 3: Checklist of the most important considerations to keep in mind during HTS data processing for detection of plant viruses

4.3.6 Which analysis approach should I choose?

The variety in similarity-based annotation approaches is striking. Choosing the most relevant one will depend on criteria such as the aims of the study (diagnostic, metagenomics) and the time/computational power available. Whichever program/approach is selected, it is important to consider its limitations and to properly set the key parameters to avoid false-positive or false-negative results. Fast programs can be used as a filtering step and then validated by slower approaches, or alternatively, two approaches can be used to validate each other. If computational time or power is not a serious limitation, combining several approaches may enhance the ability to obtain an accurate annotation [99]. We also provide a checklist, identifying the most important considerations, which should be taken into account when analyzing HTS data (Figure 3).

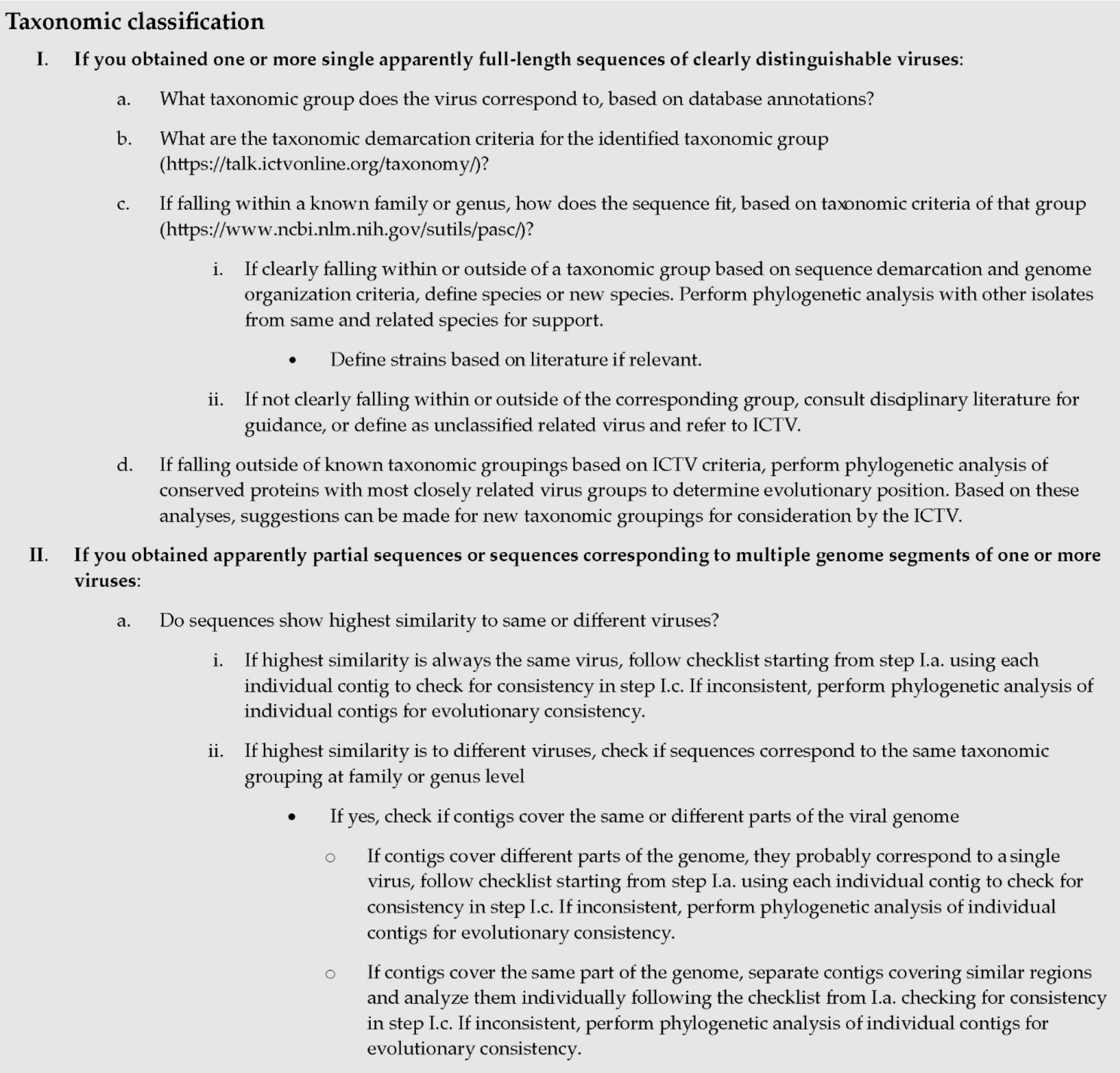
Moreover, when analyzing the data obtained from long-read technologies, one should pay special attention to using approaches which enable efficient processing of such data. Mapping algorithms have been developed for processing of long read data with higher error rates, such as Minimap2 [100]. For BLASTx-like similarity searches, algorithms, which can handle frame-shift mutations (caused by the relatively higher error rates), such as DIAMOND [87], are preferred. Assembly and polishing of long read data can improve further processing [101] and improve the chances for correct identification of viral sequences in the data.

4.3.7 Taxonomic classification

To assign viruses to taxonomic ranks species demarcation criteria specifically set for different viral genera need to be followed. Often, identities <75% at the nucleotide or protein level are indicative of a new viral species, however, the threshold might be also lower or higher, such as at <91% for begomoviruses. Identities <60% might be indicative of a new viral genus, however, the threshold might be also lower or higher, such as <45% within *Betaflexiviridae* family. As noted, these criteria differ substantially between virus families and genera but up-to-date information is published by the International Committee on Taxonomy of Viruses (ICTV) in the latest taxonomy reports [102,103] that can be found online (https://talk.ictvonline.org/taxonomy/). Once a sequence is identified to a family or genus level, a pairwise sequence comparison (PASC) webtool [104] to support virus classification, hosted by NCBI (https://www.ncbi.nlm.nih.gov/sutils/pasc/), can quickly provide an indication on how a new sequence fits in that genus or family. In cases where virus sequence identity is near the limit of the identity cut-off values for different species, additional information and/or justification may be required for their definite classification. These could include biological information such as host species, vector species or symptom types, but if enough isolates have been sequenced population genomics approaches can also be employed [105].

Strains of viruses do not fall under official taxonomy. Rather, they are definitions utilized by communities of practice around virus species and would thus require a review of the literature concerning the specific virus species to be able to classify the sequence to a particular strain or phylotype. This is a process that generally includes phylogenetic analysis of the identified sequence with published virus (reference) sequences.

The approach described above can be rather straightforward if complete genomes of viruses with a single genome segment have been assembled. However, things can become more ambiguous in situations where a new virus has multiple genome segments or have been incompletely assembled, resulting in several contigs corresponding to different parts of a viral genome. The individual contigs for a novel virus may be equally distantly related to several known viruses and can then show the highest level of similarity with different viruses, which could lead to the erroneous interpretation that several new viruses are found in the same sample. This issue will often manifest itself in the previous step of similarity searches, and to resolve this the first recommended step is to identify the taxonomic position of all the best hits identified for the different viral contigs. If several best hits fall within the same genus or family, one could suspect they may correspond to the same virus. The next step would be to investigate the general viral genome structures in the identified genus or family from the ICTV reports and ascertain if the different best hits correspond to the same or different genomic regions for that type of virus. If they are all different, it is likely that a single new species is present, if the same region is covered by multiple contigs which differ significantly from each other, then the scenario of multiple new viruses belonging to a similar taxonomic group is more probable. A checklist in Figure 4 contains most important points to keep in mind for taxonomic classification of viral sequences obtained by HTS.



Taxonomic classification

I. **If you obtained one or more single apparently full-length sequences of clearly distinguishable viruses**:

   a. What taxonomic group does the virus correspond to, based on database annotations?

   b. What are the taxonomic demarcation criteria for the identified taxonomic group (https://talk.ictvonline.org/taxonomy/)?

   c. If falling within a known family or genus, how does the sequence fit, based on taxonomic criteria of that group (https://www.ncbi.nlm.nih.gov/sutils/pasc/)?

      i. If clearly falling within or outside of a taxonomic group based on sequence demarcation and genome organization criteria, define species or new species. Perform phylogenetic analysis with other isolates from same and related species for support.

         - Define strains based on literature if relevant.

      ii. If not clearly falling within or outside of the corresponding group, consult disciplinary literature for guidance, or define as unclassified related virus and refer to ICTV.

   d. If falling outside of known taxonomic groupings based on ICTV criteria, perform phylogenetic analysis of conserved proteins with most closely related virus groups to determine evolutionary position. Based on these analyses, suggestions can be made for new taxonomic groupings for consideration by the ICTV.

II. **If you obtained apparently partial sequences or sequences corresponding to multiple genome segments of one or more viruses**:

   a. Do sequences show highest similarity to same or different viruses?

      i. If highest similarity is always the same virus, follow checklist starting from step I.a. using each individual contig to check for consistency in step I.c. If inconsistent, perform phylogenetic analysis of individual contigs for evolutionary consistency.

      ii. If highest similarity is to different viruses, check if sequences correspond to the same taxonomic grouping at family or genus level

         - If yes, check if contigs cover the same or different parts of the viral genome

            ○ If contigs cover different parts of the genome, they probably correspond to a single virus, follow checklist starting from step I.a. using each individual contig to check for consistency in step I.c. If inconsistent, perform phylogenetic analysis of individual contigs for evolutionary consistency.

            ○ If contigs cover the same part of the genome, separate contigs covering similar regions and analyze them individually following the checklist from I.a. checking for consistency in step I.c. If inconsistent, perform phylogenetic analysis of individual contigs for evolutionary consistency.

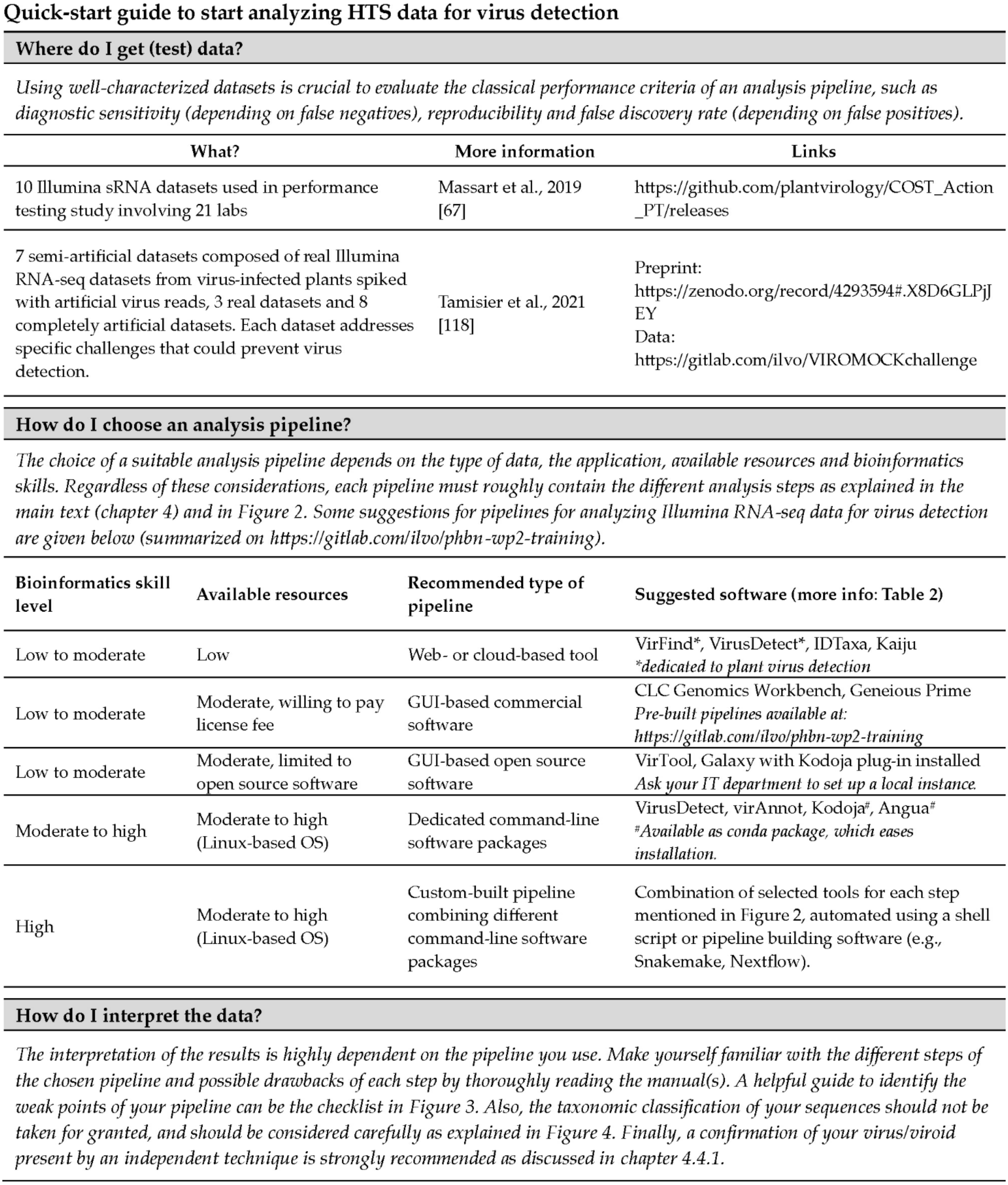
Figure 4: Checklist of the most important considerations during taxonomic classification of plant viruses detected by HTS.

New viruses belonging to previously undescribed families and/or genera can often only be reliably aligned by using the translated amino acid sequences of conserved genes such as polymerases and coat proteins. In these cases, phylogenies generated with viruses from related genera or families are needed to determine the exact taxonomic position. Additional criteria, such as number of open reading frames and overall genomic organization need to be considered to classify a virus as a member of a new genus or family. When there is uncertainty, viruses can be categorized as unclassified new species, until new evidence arises that can support a definite classification.

Irrespective of the situation encountered, to become officially recognized species, generally a near complete genome sequence, including the complete coding sequence information, is required by the ICTV to assign a 'sequence only' virus to a species level. If relevant supportive biological data is available that rule is more relaxed and will be determined by the relevant virus family study groups.

4.3.8 “Quick start” methods

Depending on the computational background of the user, there are different ways to approach the analysis. Many software solutions are available for detecting the presence of (plant) viruses in HTS datasets, summarized recently by several reviews [106,107]. For beginners or newcomers in the field, all these tools can be overwhelming. The quick-start guide (Figure 5) might be handy to select an appropriate tool or pipeline.



Quick-start guide to start analyzing HTS data for virus detection

Where do I get (test) data?

Using well-characterized datasets is crucial to evaluate the classical performance criteria of an analysis pipeline, such as diagnostic sensitivity (depending on false negatives), reproducibility and false discovery rate (depending on false positives).

What?	More information	Links
10 Illumina sRNA datasets used in performance testing study involving 21 labs	Massart et al., 2019 [67]	https://github.com/plantvirology/COST_Action_PT/releases
7 semi-artificial datasets composed of real Illumina RNA-seq datasets from virus-infected plants spiked with artificial virus reads, 3 real datasets and 8 completely artificial datasets. Each dataset addresses specific challenges that could prevent virus detection.	Tamisier et al., 2021 [118]	Preprint: https://zenodo.org/record/4293594#.X8D6GLPjJEY Data: https://gitlab.com/ilvo/VIROMOCKchallenge

How do I choose an analysis pipeline?

The choice of a suitable analysis pipeline depends on the type of data, the application, available resources and bioinformatics skills. Regardless of these considerations, each pipeline must roughly contain the different analysis steps as explained in the main text (chapter 4) and in Figure 2. Some suggestions for pipelines for analyzing Illumina RNA-seq data for virus detection are given below (summarized on https://gitlab.com/ilvo/phbn-wp2-training).

Bioinformatics skill level	Available resources	Recommended type of pipeline	Suggested software (more info: Table 2)
Low to moderate	Low	Web- or cloud-based tool	VirFind*, VirusDetect*, IDTaxa, Kaiju **dedicated to plant virus detection*
Low to moderate	Moderate, willing to pay license fee	GUI-based commercial software	CLC Genomics Workbench, Geneious Prime *Pre-built pipelines available at: https://gitlab.com/ilvo/phbn-wp2-training*
Low to moderate	Moderate, limited to open source software	GUI-based open source software	VirTool, Galaxy with Kodoja plug-in installed *Ask your IT department to set up a local instance.*
Moderate to high	Moderate to high (Linux-based OS)	Dedicated command-line software packages	VirusDetect, virAnnot, Kodoja#, Angua# *#Available as conda package, which eases installation.*
High	Moderate to high (Linux-based OS)	Custom-built pipeline combining different command-line software packages	Combination of selected tools for each step mentioned in Figure 2, automated using a shell script or pipeline building software (e.g., Snakemake, Nextflow).

How do I interpret the data?

The interpretation of the results is highly dependent on the pipeline you use. Make yourself familiar with the different steps of the chosen pipeline and possible drawbacks of each step by thoroughly reading the manual(s). A helpful guide to identify the weak points of your pipeline can be the checklist in Figure 3. Also, the taxonomic classification of your sequences should not be taken for granted, and should be considered carefully as explained in Figure 4. Finally, a confirmation of your virus/viroid present by an independent technique is strongly recommended as discussed in chapter 4.4.1.

Figure 5: Quick-start guide assisting selection of analysis approaches for plant virus detection from HTS data.

Among these options, easy-to-use pipelines that do not requiring extensive computational expertise might be a good start. These pipelines present a user-friendly interface on-line or directly on the computer. A first group of pipelines can be considered as "all in one": they automatically start on the raw data to deliver the final results as a list of viruses detected. They may or may not allow the adaptation of parameters. A second group corresponds to pipelines for which the different steps of the process have to be done separately and independently. This is the case when using commercial software such as CLC Genomics Workbench or Geneious Prime, which both also enable the building of customized "all-in-one" workflows. Table A1 summarizes the pros and cons of the most common "easy-to-use" analysis solutions. Ease of use may generate a false sense of confidence in the results and as with all pipelines. Understanding of the steps and the parameters of the pipelines, as well as critical interpretation of the results is always required.

4.4 What to do when the data analysis is concluded?

4.4.1 Identity confirmation by an independent technique

As for many other test methods, HTS may sometimes provide false positive results. If consequential, it is therefore important that HTS results are confirmed.

The need to confirm the identity of a pest depends on the context of the analysis and on the type of organism identified (e.g., identification of a quarantine compared to an endemic pest). The results must be confirmed in cases considered critical to national or international plant protection programs. These are the detection of a pest in an area where it is not known to occur or in a consignment originating from a country where it is declared to be absent; and also, when a pest is identified by a laboratory for the first time (EPPO PM 7/76, 2019). The identity of any uncharacterized pest with potential risks to plant health should also be confirmed by another test. Whilst a virus in its common host is unlikely to require confirmation (if not regulated), it may be useful if associated with different symptoms (e.g., an emerging strain) or if detected in a new host.

When confirmation is needed, it is recommended to use a test or a combination of tests based on different biological principles (e.g. ELISA or targeted PCR instead of resequencing the sample using the same protocol). If available, validated tests should be used and a new sample extract obtained for analysis. The selection of confirmatory tests depends on the performance characteristics required, the general characteristics of methods for plant virology have been reviewed by Roenhorst et al. (2018). If no other tests are available to confirm the identity of the pest (i.e., poorly characterized and uncharacterized organisms), primers should be designed and tested, based on the HTS sequence data and available sequence information in the sequence databases. Alternatively, generic primers that enable the amplification of viruses within a genus or family, including the targeted one(s), followed by Sanger sequencing of the amplicons could be used to confirm the identity.

4.4.2 Biological characterization post HTS detection

Based on HTS, the list of thus far unknown or poorly characterized viruses for which only genome data are available is rapidly increasing [109]. This presents a challenge for the further steps necessary to determine the causative relationship to a disease and guide phytosanitary diagnostic laboratories on data interpretation and recommendations. Viruses for which only genome data are available can indeed be taxonomically assigned but the real challenge is to attribute biological meaning to their detection. The interpretation of the biological relevance applies mainly to poorly characterized and uncharacterized or newly discovered viruses. For example, the viral sequences detected may correspond to a *bona fide* virus infecting other organisms associated with the sample, including bacteria, fungi or arthropods [110,111] or to viral sequences integrated into the plant genome [112,113]. As stated previously [113], relevant scientific expertise is essential for sound biological interpretation of HTS results, in particular when identifying a target with a low titer, a poorly characterized species, an uncharacterized organism or sequences integrated in the host genome [67,114].

The extent to which additional biological characterization is performed depends largely on the potential risk the organism(s) would pose to plant health, although acquisition of such data may take time or may not be possible (e.g., lack of human and/or financial resources). The scaled and progressive scientific framework proposed by Massart et al. (2017) is a useful tool for guiding the biological characterization and the risk assessment of an uncharacterized or poorly characterized plant virus detected by HTS.

4.4.3 Sharing data to leverage knowledge

After the detection of the virus in the laboratory, the researcher or diagnostician faces an important dilemma: when and how to share data publicly. As shown by recent examples [115–117], pre-publication data sharing between laboratories brings valuable information to address the risks raised by a virus. Sharing data will give a more global picture of its geographical repartition, its genetic diversity, its host range and symptomatology, allowing a contextualized risk analysis and avoiding unnecessary regulatory action. When shared, the genome information usefulness is leveraged. Data sharing must also include metadata from the sample (e.g., origin, species, cultivar, time point, organ of sampling). Nevertheless, data sharing is not always easy due to regulatory implications and for commercial work laboratories may be bound by confidentiality agreements [7]. Besides sharing sequence data itself, sharing of analysis pipelines, protocols and experiences between labs can greatly contribute to the harmonization of the field and provide useful resources for newcomers to the field. The recently established Plant Health Bioinformatics Network (PHBN) aims to foster this approach and provide protocols, pipelines (https://gitlab.com/ilvo/phbn-wp2-training) and reference datasets (https://gitlab.com/ilvo/VIROMOCKchallenge) [118] that can be widely employed. It also aims to organize community efforts to advance certain aspects of plant health bioinformatics (https://gitlab.com/ilvo/PHBN-WP4-RNAseq_Community_Screening).

4.4.4 Additional bioinformatics analyses

Further analyses, beyond viral detection and taxonomic classification, can be performed on HTS data, depending on the goal of the study. For instance, the large amount of sequence generated by HTS allows a good resolution of the within-host genetic diversity of the viral populations. Assessing the genetic diversity within and among viral populations can provide a better understanding of virus evolution and help to determine population genetic parameters or epidemiological patterns. This can be done using single nucleotide polymorphism (SNP) calling algorithms, which need to allow detection of low frequency variants expected in virus populations. Other analysis, like genetic recombination detection, can also be performed. The most popular software solutions, which detect recombination patterns comparing full or partial viral genomes and run on Windows, are RDP4 [119], SimPlot [120] and TreeOrder Scan [121]. ViReMa (Viral Recombination Mapper) can be used for detection of recombination junctions, as well as insertion/substitution events and multiple recombinations within single reads [122], and has been successfully applied for the analysis of recombination events in plant virus genomes [22,123,124]. Phylogenetic relationships among the detected and previously known viruses can also be investigated using fast neighbor-joining algorithms [125], more precise maximum likelihood approaches [126,127] or Bayesian analysis approaches [128]. Freeware phylogenetic analysis suites, such as MEGAN [129], or phylogenetic analysis algorithms integrated within commercial software, such as CLC Genomics Workbench and Geneious Prime, can be used. Studying the time of emergence of viral species and strains including the distribution of the genetic diversity across geographical sites can be done using software such as BEAST [130] and SPAGeDI [131].

5. Conclusions and outlook

In this review we aimed to provide an informative primer on the generation and analysis of HTS data for detection of plant viruses. Even though the field of HTS is transforming rapidly and new platforms and analysis tools are being developed constantly, the basic concepts of data analysis reviewed here will remain relevant in the future. In the next few years, we expect a great increase in the use of the long read HTS platforms. New algorithms and pipelines for analysis of data will continually be developed, building on some of the concepts described above. These developments are likely to focus in two main areas. Firstly, the adoption of deep learning approaches will likely be more and more integrated into the field of virus detection, on different levels, from similarity searches to the estimation of detection confidence levels, to enable the more robust detection of virus sequences that are more distantly related to those we currently recognize. Secondly, with the further development of nanopore sequencing-based platforms, potentially facilitating on-site HTS analysis of samples, we will need faster and more memory-efficient analysis approaches to enable rapid data analysis, potentially away from centralized facilities. Moreover, guidelines are being developed to enable validation and verification of HTS-based detection of plant pathogens in research and diagnostic settings, which also include bioinformatics steps of the analysis [9]. These guidelines will provide detailed information on how to use appropriate controls and which specific results parameters to use to ensure the validity of the results, briefly covered in Figure 3 and Figure 4 in this text. Finally, we encourage the readers to use this guide as a starting point for the selection of appropriate analysis approaches and to get further informed about the specifics of the algorithms (Figure 5). By combining knowledge on the analysis approaches with a sound plant virology background, we can maximize the potential of these technologies and provide sound interpretation of the results.

Author Contributions: Conceptualization, D.K., L.T., S.M. and A.H.; writing—original draft preparation, all authors; writing—review and editing, all authors. All authors have read and agreed to the published version of the manuscript.

Funding: This review was partially funded by the Belgian Federal Public Service of Health, Food Chain Safety and Environment (FPS Health) through the contract "RI 18_A-289", Slovenian Research Agency (core funding P4-0407 and project L7-2632) and by the Euphresco project "Plant Health Bioinformatics Network" (PHBN) (2018-A-289).

Acknowledgments: We thank Olivera Maksimović Carvalho Ferreira, Nuria Fontdevila, Maryam Khalili, Ayoub Maachi, Mark Paul Selda Rivarez and Deborah Schönegger for reading and commenting an earlier draft of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

770 ## Appendix

771 **Table A1:** List of selected easy-to-use analysis solutions for detection of plant viruses with their pros and cons

Pipeline	Brief description	Web link / Publication	Pros	Cons
Virusdetect	Virus discovery using sRNA and RNAseq sequences	http://virusdetect.feilab.net [132]	• Easy to use: single command to run one or multiple datasets simultaneously. • Performs *de novo* assembly and reference mapping in parallel, including optional host genome subtraction and identified contigs through BLASTn & BLASTx. • Automatic results organization and presentation in html table providing key metrics on coverage, sequence depth, virus and genus name and link to visual map and NCBI GenBank reference sequence. • Options to modify key assembly, mapping and reporting parameters. • Windows version with visual interface & automatic quality control and trimming to be released in 2021. • Available via user account online.	• Uses complete NCBI GenBank database for viruses (divided along host type) for reference mapping and identity searches. NCBI GenBank sequences are poorly curated and may lead to reports of wrong results. • Creating & formatting new custom or up to date NCBI GenBank reference library is not very straightforward and ready formatted updates are not uploaded very regularly to the VirusDetect webpage. • Currently requires Linux environment, which is an impediment for many diagnosticians. • Default reporting cutoff settings are optimized for siRNA to minimize false positives due to index-hopping, however may lead to non-reporting of low concentration viruses.
Virtool	HTS sample manager with virus detection, discovery and analysis workflows	www.virtool.ca https://github.com/virtool/virtool [36]	• Open source modern graphical optimized for cloud computing. • User and group control with password protection; sample data management; security and QA features. • Support for multiple workflows and versioned databases for viral and non-viral pathogens. • Can process short and long reads (Illumina, Oxford Nanopore Technologies). • Result visualization, filtering, and sorting. • HTTP API for automation or integration with other services such as LIMS.	• Requires some more computational skills for user (or help of informatician) to install a local server on Linux operating system. • limited ability to change parameters within a workflow.

Pipeline	Brief description	Web link / Publication	Pros	Cons
virAnnot	Command-line tool for virus detection and viral diversity estimation	[133]	• Wide options to modify assembly, mapping, annotation and clustering parameters. • Performs parallel analysis of samples from the same dataset. • Estimation of viral diversity through Operational Taxonomic Units (OTUs). • Easy results visualization with Krona and phylogenic trees.	• Requires a Linux environment, which is an impediment for many diagnosticians. • Need a cluster access for the annotation step. • Requires a good knowledge of command-line and Unix packages installation.
VirFind	Online virus discovery tool	http://virfind.org [134]	• Available via user account online. • Performs reference mapping, *de novo* assembly and conserved domain searches in parallel or subsequently.	• Analysis by online version can take several days. • Output only in text files: experience needed for further interpretation.
Angua	Command-line tool for virus detection	https://fred.fera.co.uk/smc-greig/angua3	• Simple - can be executed with one command, but has a number of parameters/tools which can be tweaked • Uses full nt and nr GenBank databases so is sensitive • Manual inspection of results with a local MEGAN installation improves accuracy • Supports single and paired-end analysis • Supports blastn/MEGAN parallelisation	• Requires a Linux environment, which is an impediment for many diagnosticians. • Dependent on locally stored nt and nr GenBank databases. • Blastx stage can take a long time. • Manual inspection of results with a local MEGAN installation is required.
Kodoja	k-mer based command-line tool for virus detection	https://github.com/abaizan/kodoja [95]	• Available as Galaxy plug-in or as command-line tool that can be installed using conda. • k-mer based rather than assembly and mapping, which makes it more sensitive and computationally less intensive.	• Requires a Linux environment for the command-line tool, which is an impediment for many diagnosticians.
Truffle	Targeted virus detection using e-probes based approach	[135]	• Results easy to interpret, good sensitivity. • Requires relatively low computational resources.	• Undescribed virus or viral strain will not be detectable using this pipeline. • Only grapevine and citrus viruses are available, however e-probes for other viruses can be designed. • Requires a Linux environment, which is an impediment for many diagnosticians.
Kaiju	Online metagenomic analysis tool	http://kaiju.binf.ku.dk/ [93]	• Both standalone and web server available. • Quick analysis not requiring any knowledge in bioinformatics and data analysis. • Prepared downloadable databases available.	• Not specifically made for virus detection. • Protein based, hence blind for non-coding sequences (viroids, satellites).

Pipeline	Brief description	Web link / Publication	Pros	Cons
Galaxy	Workflow system for computational analyses	https://usegalaxy.org [136]	• Web-based platform. • Open source. • Vast choice of computational biology tools.	• Limit in data upload, unless if you establish own local galaxy server. • Not specifically made for virus detection.
ID-Seq	Online metagenomic analysis tool	https://id-seq.net/ [137]	• Easy-to-use visual interface of results. • Quick analysis not requiring any knowledge in bioinformatics and data analysis.	• Not possible to change parameters of the workflow. • Complementary software needed for reads alignment. • Not specifically made for virus detection.
Geneious Prime	Software for molecular biology and sequence analysis	https://www.geneious.com	• Graphical interface. • Multiple plugins available, including some frequently used freeware assembly algorithms. • Automated, customizable workflows. • Constant release of updated versions and customer support. • Nice and efficient visualization tools. • Free trial version available.	• Licensed, including license fee; • HTS data analysis requires computational resources.
CLC Genomics Workbench	Comprehensive software solution of molecular biology analysis tools	https://digitalinsights.qiagen.com/products-overview/discovery-insights-portfolio/analysis-and-visualization/qiagen-clc-genomics-workbench/	• Graphical interface. • Automated, customizable workflows. • Constant release of updated versions and customer support. • Nice and efficient visualization tools. • Free trial version available.	• Expensive ongoing licensing fee. • HTS data analysis requires computational resources.

772

773

References

1. Villamor, D.E.V.; Ho, T.; Al Rwahnih, M.; Martin, R.R.; Tzanetakis, I.E. High throughput sequencing for plant virus detection and discovery. *Phytopathology* **2019**, *109*, 716–725, doi:10.1094/PHYTO-07-18-0257-RVW.
2. Kreuze, J.F.; Perez, A.; Untiveros, M.; Quispe, D.; Fuentes, S.; Barker, I.; Simon, R. Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: A generic method for diagnosis , discovery and sequencing of viruses. *Virology* **2009**, *388*, 1–7, doi:10.1016/j.virol.2009.03.024.
3. Adams, I.P.; Glover, R.H.; Monger, W.A.; Mumford, R.; Jackeviciene, E.; Navalinskiene, M.; Samuitiene, M.; Boonham, N. Next-generation sequencing and metagenomic analysis: A universal diagnostic tool in plant virology. *Mol. Plant Pathol.* **2009**, *10*, 537–545, doi:10.1111/j.1364-3703.2009.00545.x.
4. Al Rwahnih, M.; Daubert, S.; Golino, D.; Rowhani, A. Deep sequencing analysis of RNAs from a grapevine showing Syrah decline symptoms reveals a multiple virus infection that includes a novel virus. *Virology* **2009**, *387*, 395–401, doi:10.1016/j.virol.2009.02.028.
5. Donaire, L.; Wang, Y.; Gonzalez-Ibeas, D.; Mayer, K.F.; Aranda, M.A.; Llave, C. Deep-sequencing of plant viral small RNAs reveals effective and widespread targeting of viral genomes. *Virology* **2009**, *392*, 203–214, doi:10.1016/j.virol.2009.07.005.
6. Massart, S.; Chiumenti, M.; De Jonghe, K.; Glover, R.; Haegeman, A.; Koloniuk, I.; Kominek, P.; Kreuze, J.; Kutnjak, D.; Lotos, L.; et al. Virus detection by high-throughput sequencing of small RNAs: large scale performance testing of sequence analysis strategies. *Phytopathology* **2018**, doi:10.1094/PHYTO-02-18-0067-R.
7. Olmos, A.; Boonham, N.; Candresse, T.; Gentit, P.; Giovani, B.; Kutnjak, D.; Liefting, L.; Maree, H.J.; Minafra, A.; Moreira, A.; et al. High-throughput sequencing technologies for plant pest diagnosis: challenges and opportunities. *EPPO Bull.* **2018**, *48*, 219–224, doi:10.1111/epp.12472.
8. Weymann, D.; Laskin, J.; Roscoe, R.; Schrader, K.A.; Chia, S.; Yip, S.; Cheung, W.Y.; Gelmon, K.A.; Karsan, A.; Renouf, D.J.; et al. The cost and cost trajectory of whole-genome analysis guiding treatment of patients with advanced cancers. *Mol. Genet. Genomic Med.* **2017**, *5*, 251–260, doi:10.1002/mgg3.281.
9. Valitest EU project consortium *Guidelines for the selection, development, validation and routine use of high-throughput sequencing analysis in plant health diagnostic laboratories: grant agreement N. 773139: deliverable N° 2.2. (confidential)*; 2020;
10. Maliogka, V.I.; Minafra, A.; Saldarelli, P.; Ruiz-García, A.B.; Glasa, M.; Katis, N.; Olmos, A. Recent advances on detection and characterization of fruit tree viruses using high-throughput sequencing technologies. *Viruses* **2018**, *10*, 436, doi:10.3390/v10080436.
11. Roossinck, M.J. Deep sequencing for discovery and evolutionary analysis of plant viruses. *Virus Res.* **2017**, *239*, 82–86, doi:10.1016/j.virusres.2016.11.019.
12. Roossinck, M.J.; Martin, D.P.; Roumagnac, P. Plant virus metagenomics: Advances in virus discovery. *Phytopathology* **2015**, *105*, 716–727, doi:10.1094/PHYTO-12-14-0356-RVW.
13. Marais, A.; Faure, C.; Bergey, B.; Candresse, T. Viral Double-Stranded RNAs (dsRNAs) from Plants: Alternative Nucleic Acid Substrates for High-Throughput Sequencing. In *Viral Metagenomics: Methods and Protocols, Methods in Molecular Biology, vol. 1746*; Pantaleo, V., Chiumenti, M., Eds.; Springer Nature: New York, 2018; pp. 45–53 ISBN 978-1-4939-7682-9.
14. Massart, S.; Olmos, A.; Jijakli, H.; Candresse, T. Current impact and future directions of high throughput sequencing in plant virus diagnostics. *Virus Res.* **2014**, *188*, 90–96, doi:10.1016/j.virusres.2014.03.029.
15. Pecman, A.; Kutnjak, D.; Gutiérrez-Aguirre, I.; Adams, I.; Fox, A.; Boonham, N.; Ravnikar, M. Next generation sequencing for detection and discovery of plant viruses and viroids: Comparison of two approaches. *Front. Microbiol.* **2017**, *8*, doi:10.3389/fmicb.2017.01998.
16. Boone, M.; De Koker, A.; Callewaert, N. Survey and summary capturing the "ome": The expanding molecular toolbox for RNA and DNA library construction. *Nucleic Acids Res.* **2018**, *46*, 2701–2721, doi:10.1093/nar/gky167.

17. Visser, M.; Bester, R.; Burger, J.T.; Maree, H.J. Next-generation sequencing for virus detection: Covering all the bases. *Virol. J.* **2016**, *13*, 4–9, doi:10.1186/s12985-016-0539-x.
18. Idris, A.; Al-Saleh, M.; Piatek, M.J.; Al-Shahwan, I.; Ali, S.; Brown, J.K. Viral metagenomics: Analysis of begomoviruses by illumina high-throughput sequencing. *Viruses* **2014**, *6*, 1219–1236, doi:10.3390/v6031219.
19. Sukal, A.C.; Kidanemariam, D.B.; Dale, J.L.; Harding, R.M.; James, A.P. Assessment and optimization of rolling circle amplification protocols for the detection and characterization of badnaviruses. *Virology* **2019**, *529*, 73–80, doi:10.1016/j.virol.2019.01.013.
20. Wyant, P.S.; Strohmeier, S.; Schäfer, B.; Krenz, B.; Assunção, I.P.; Lima, G.S. de A.; Jeske, H. Circular DNA genomics (circomics) exemplified for geminiviruses in bean crops and weeds of northeastern Brazil. *Virology* **2012**, *427*, 151–157, doi:10.1016/j.virol.2012.02.007.
21. Vivek, A.T.; Zahra, S.; Kumar, S. From current knowledge to best practice: A primer on viral diagnostics using deep sequencing of virus-derived small interfering RNAs (vsiRNAs) in infected plants. *Methods* **2019**, doi:10.1016/j.ymeth.2019.10.009.
22. Kutnjak, D.; Rupar, M.; Gutierrez-Aguirre, I.; Curk, T.; Kreuze, J.F.; Ravnikar, M. Deep Sequencing of Virus-Derived Small Interfering RNAs and RNA from Viral Particles Shows Highly Similar Mutational Landscapes of a Plant Virus Population. *J. Virol.* **2015**, *89*, 4760–4769, doi:10.1128/JVI.03685-14.
23. Seguin, J.; Rajeswaran, R.; Malpica-López, N.; Martin, R.R.; Kasschau, K.; Dolja, V. V.; Otten, P.; Farinelli, L.; Pooggin, M.M. De novo reconstruction of consensus master genomes of plant RNA and DNA viruses from siRNAs. *PLoS One* **2014**, *9*, 1–8, doi:10.1371/journal.pone.0088513.
24. Smith, O.; Clapham, A.; Rose, P.; Liu, Y.; Wang, J.; Allaby, R.G. A complete ancient RNA genome: Identification, reconstruction and evolutionary history of archaeological Barley Stripe Mosaic Virus. *Sci. Rep.* **2014**, *4*, 1–6, doi:10.1038/srep04003.
25. Turco, S.; Golyaev, V.; Seguin, J.; Gilli, C.; Farinelli, L.; Boller, T.; Schumpp, O.; Pooggin, M.M. Small RNA-omics for virome reconstruction and antiviral defense characterization in mixed infections of cultivated solanum plants. *Mol. Plant-Microbe Interact.* **2018**, *31*, 707–723, doi:10.1094/MPMI-12-17-0301-R.
26. Melcher, U.; Muthukumar, V.; Wiley, G.B.; Min, B.E.; Palmer, M.W.; Verchot-Lubicz, J.; Ali, A.; Nelson, R.S.; Roe, B.A.; Thapa, V.; et al. Evidence for novel viruses by analysis of nucleic acids in virus-like particle fractions from Ambrosia psilostachya. *J. Virol. Methods* **2008**, *152*, 49–55, doi:10.1016/j.jviromet.2008.05.030.
27. Muthukumar, V.; Melcher, U.; Pierce, M.; Wiley, G.B.; Roe, B.A.; Palmer, M.W.; Thapa, V.; Ali, A.; Ding, T. Non-cultivated plants of the Tallgrass Prairie Preserve of northeastern Oklahoma frequently contain virus-like sequences in particulate fractions. *Virus Res.* **2009**, *141*, 169–173, doi:10.1016/j.virusres.2008.06.016.
28. Bernardo, P.; Charles-Dominique, T.; Barakat, M.; Ortet, P.; Fernandez, E.; Filloux, D.; Hartnady, P.; Rebelo, T.A.; Cousins, S.R.; Mesleard, F.; et al. Geometagenomics illuminates the impact of agriculture on the distribution and prevalence of plant viruses at the ecosystem scale. *ISME J.* **2018**, *12*, 173–184, doi:10.1038/ismej.2017.155.
29. Filloux, D.; Dallot, S.; Delaunay, A.; Galzi, S.; Jacquot, E.; Roumagnac, P. Metagenomics approaches based on virion-associated nucleic acids (VANA): An innovative tool for assessing without a priori viral diversity of plants. *Methods Mol. Biol.* **2015**, *1302*, 249–257, doi:10.1007/978-1-4939-2620-6_18.
30. Ma, Y.; Marais, A.; Lefebvre, M.; Theil, S.; Svanella-Dumas, L.; Faure, C.; Candresse, T. Phytovirome Analysis of Wild Plant Populations: Comparison of Double-Stranded RNA and Virion-Associated Nucleic Acid Metagenomic Approaches. *J. Virol.* **2019**, *94*, doi:10.1128/jvi.01462-19.
31. Roossinck, M.J. Plants, viruses and the environment: Ecology and mutualism. *Virology* **2015**, *479–480*, 271–277, doi:10.1016/j.virol.2015.03.041.

32. Hull, R. Origins and Evolution of Plant Viruses. In *Plant Virology*; Elsevier, 2014; pp. 423–476.
33. Al Rwahnih, M.; Daubert, S.; Golino, D.; Islas, C.; Rowhani, A. Comparison of next-generation sequencing versus biological indexing for the optimal detection of viral pathogens in grapevine. *Phytopathology* **2015**, *105*, 758–763, doi:10.1094/PHYTO-06-14-0165-R.
34. Kesanakurti, P.; Belton, M.; Saeed, H.; Rast, H.; Boyes, I.; Rott, M. Screening for plant viruses by next generation sequencing using a modified double strand RNA extraction protocol with an internal amplification control. *J. Virol. Methods* **2016**, *236*, 35–40, doi:10.1016/j.jviromet.2016.07.001.
35. Loconsole, G.; Saldarelli, P.; Doddapaneni, H.; Savino, V.; Martelli, G.P.; Saponari, M. Identification of a single-stranded DNA virus associated with citrus chlorotic dwarf disease, a new member in the family Geminiviridae. *Virology* **2012**, *432*, 162–172, doi:10.1016/j.virol.2012.06.005.
36. Rott, M.; Xiang, Y.; Boyes, I.; Belton, M.; Saeed, H.; Kesanakurti, P.; Hayes, S.; Lawrence, T.; Birch, C.; Bhagwat, B.; et al. Application of next generation sequencing for diagnostic testing of tree fruit viruses and viroids. *Plant Dis.* **2017**, *101*, 1489–1499, doi:10.1094/PDIS-03-17-0306-RE.
37. Weber, F.; Wagner, V.; Rasmussen, S.B.; Hartmann, R.; Paludan, S.R. Double-Stranded RNA Is Produced by Positive-Strand RNA Viruses and DNA Viruses but Not in Detectable Amounts by Negative-Strand RNA Viruses. *J. Virol.* **2006**, *80*, 5059–5064, doi:10.1128/jvi.80.10.5059-5064.2006.
38. Gaafar, Y.Z.A.; Ziebell, H. Comparative study on three viral enrichment approaches based on RNA extraction for plant virus/viroid detection using high-throughput sequencing. *PLoS One* **2020**, *15*, 1–17, doi:10.1371/journal.pone.0237951.
39. Thapa, V.; McGlinn, D.J.; Melcher, U.; Palmer, M.W.; Roossinck, M.J. Determinants of taxonomic composition of plant viruses at the Nature Conservancy's Tallgrass Prairie Preserve, Oklahoma. *Virus Evol.* **2015**, *1*, doi:10.1093/ve/vev007.
40. Blouin, A.G.; Ross, H.A.; Hobson-Peters, J.; O'Brien, C.A.; Warren, B.; MacDiarmid, R. A new virus discovered by immunocapture of double-stranded RNA, a rapid method for virus enrichment in metagenomic studies. *Mol. Ecol. Resour.* **2016**, *16*, 1255–1263, doi:10.1111/1755-0998.12525.
41. Kobayashi, K.; Tomita, R.; Sakamoto, M. Recombinant plant dsRNA-binding protein as an effective tool for the isolation of viral replicative form dsRNA and universal detection of RNA viruses. *J. Gen. Plant Pathol.* **2009**, *75*, 87–91, doi:10.1007/s10327-009-0155-3.
42. Roossinck, M.J.; Saha, P.; Wiley, G.B.; Quan, J.; White, J.D.; Lai, H.; Chavarría, F.; Shen, G.; Roe, B.A. Ecogenomics: Using massively parallel pyrosequencing to understand virus ecology. *Mol. Ecol.* **2010**, *19*, 81–88, doi:10.1111/j.1365-294X.2009.04470.x.
43. Chalupowicz, L.; Dombrovsky, A.; Gaba, V.; Luria, N.; Reuven, M.; Beerman, A.; Lachman, O.; Dror, O.; Nissan, G.; Manulis-Sasson, S. Diagnosis of plant diseases using the Nanopore sequencing platform. *Plant Pathol.* **2019**, *68*, 229–238, doi:10.1111/ppa.12957.
44. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **2011**, *17*, 10, doi:10.14806/ej.17.1.200.
45. Illumina bcl2fastq and bcl2fastq2 Conversion Software 2019.
46. Oxford Nanopore Technologies Guppy: Local accelerated basecalling for Nanopore data 2018.
47. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120, doi:10.1093/bioinformatics/btu170.
48. Wick, B. Porechop 2017.
49. De Coster, W.; D'Hert, S.; Schultz, D.T.; Cruts, M.; Van Broeckhoven, C. NanoPack: Visualizing and processing long-read sequencing data. *Bioinformatics* **2018**, *34*, 2666–2669, doi:10.1093/bioinformatics/bty149.
50. Cock, P.J.A.; Fields, C.J.; Goto, N.; Heuer, M.L.; Rice, P.M. The Sanger FASTQ file format for sequences with quality scores,

and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* **2009**, *38*, 1767–1771, doi:10.1093/nar/gkp1137.

51. Andrews, S. FastQC 2010.
52. Ewels, P.; Magnusson, M.; Lundin, S.; Käller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **2016**, *32*, 3047–3048, doi:10.1093/bioinformatics/btw354.
53. Loman, N.J.; Quinlan, A.R. Poretools: A toolkit for analyzing nanopore sequence data. *Bioinformatics* **2014**, *30*, 3399–3401, doi:10.1093/bioinformatics/btu555.
54. Najoshi sickle - A windowed adaptive trimming tool for FASTQ files using quality 2011.
55. Andino, R.; Domingo, E. Viral quasispecies. *Virology* **2015**, *479–480*, 46–51, doi:10.1016/j.virol.2015.03.022.
56. Paszkiewicz, K.; Studholme, D.J. De novo assembly of short sequence reads. *Brief. Bioinform.* **2010**, *11*, 457–472, doi:10.1093/bib/bbq020.
57. Sohn, J. Il; Nam, J.W. The present and future of de novo whole-genome assembly. *Brief. Bioinform.* **2018**, *19*, 23–40, doi:10.1093/bib/bbw096.
58. Luo, R.; Liu, B.; Xie, Y.; Li, Z.; Huang, W.; Yuan, J.; He, G.; Chen, Y.; Pan, Q.; Liu, Y.; et al. SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *Gigascience* **2012**, *1*, doi:10.1186/2047-217X-1-18.
59. Gnerre, S.; MacCallum, I.; Przybylski, D.; Ribeiro, F.J.; Burton, J.N.; Walker, B.J.; Sharpe, T.; Hall, G.; Shea, T.P.; Sykes, S.; et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 1513–1518, doi:10.1073/pnas.1017351108.
60. Simpson, J.T.; Wong, K.; Jackman, S.D.; Schein, J.E.; Jones, S.J.M.; Birol, I. ABySS: A parallel assembler for short read sequence data. *Genome Res.* **2009**, *19*, 1117–1123, doi:10.1101/gr.089532.108.
61. Zerbino, D.R.; Birney, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **2008**, *18*, 821–829, doi:10.1101/gr.074492.107.
62. Peng, Y.; Leung, H.C.M.; Yiu, S.M.; Chin, F.Y.L. IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **2012**, *28*, 1420–1428, doi:10.1093/bioinformatics/bts174.
63. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A.A.; Dvorkin, M.; Kulikov, A.S.; Lesin, V.M.; Nikolenko, S.I.; Pham, S.; Prjibelski, A.D.; et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **2012**, *19*, 455–477, doi:10.1089/cmb.2012.0021.
64. Nurk, S.; Bankevich, A.; Antipov, D.; Gurevich, A.A.; Korobeynikov, A.; Lapidus, A.; Prjibelski, A.D.; Pyshkin, A.; Sirotkin, A.; Sirotkin, Y.; et al. Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J. Comput. Biol.* **2013**, *20*, 714–737, doi:10.1089/cmb.2013.0084.
65. Bushmanova, E.; Antipov, D.; Lapidus, A.; Prjibelski, A.D. RnaSPAdes: A de novo transcriptome assembler and its application to RNA-Seq data. *Gigascience* **2019**, *8*, 1–13, doi:10.1093/gigascience/giz100.
66. Edwards, D.J.; Holt, K.E. Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. *Microb. Inform. Exp.* **2013**, *3*, doi:10.1186/2042-5783-3-2.
67. Massart, S.; Chiumenti, M.; De Jonghe, K.; Glover, R.; Haegeman, A.; Koloniuk, I.; Komínek, P.; Kreuze, J.; Kutnjak, D.; Lotos, L.; et al. Virus detection by high-throughput sequencing of small RNAs: Large-scale performance testing of sequence analysis strategies. *Phytopathology* **2019**, *109*, 488–497, doi:10.1094/PHYTO-02-18-0067-R.
68. Rang, F.J.; Kloosterman, W.P.; de Ridder, J. From squiggle to basepair: Computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.* **2018**, *19*, 1–11, doi:10.1186/s13059-018-1462-9.
69. Koren, S.; Schatz, M.C.; Walenz, B.P.; Martin, J.; Howard, J.T.; Ganapathy, G.; Wang, Z.; Rasko, D.A.; McCombie, W.R.; Jarvis, E.D.; et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **2012**, *30*, 693–700, doi:10.1038/nbt.2280.
70. Koren, S.; Walenz, B.P.; Berlin, K.; Miller, J.R.; Bergman, N.H.; Phillippy, A.M. Canu: Scalable and accurate long-read

assembly via adaptive κ-mer weighting and repeat separation. *Genome Res.* **2017**, *27*, 722–736, doi:10.1101/gr.215087.116.

71. Chin, C.S.; Peluso, P.; Sedlazeck, F.J.; Nattestad, M.; Concepcion, G.T.; Clum, A.; Dunn, C.; O'Malley, R.; Figueroa-Balderas, R.; Morales-Cruz, A.; et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **2016**, *13*, 1050–1054, doi:10.1038/nmeth.4035.
72. Oxford Nanopore Technologies Pomoxis - bioinformatics tools for nanopore research 2018.
73. Filloux, D.; Fernandez, E.; Loire, E.; Claude, L.; Galzi, S.; Candresse, T.; Winter, S.; Jeeva, M.L.; Makeshkumar, T.; Martin, D.P.; et al. Nanopore-based detection and characterization of yam viruses. *Sci. Rep.* **2018**, *8*, doi:10.1038/s41598-018-36042-7.
74. Boykin, L.M.; Sseruwagi, P.; Alicai, T.; Ateka, E.; Mohammed, I.U.; Stanton, J.A.L.; Kayuki, C.; Mark, D.; Fute, T.; Erasto, J.; et al. Tree lab: Portable genomics for early detection of plant viruses and pests in sub-saharan africa. *Genes (Basel).* **2019**, *10*, 632, doi:10.3390/genes10090632.
75. Naito, F.Y.B.; Melo, F.L.; Fonseca, M.E.N.; Santos, C.A.F.; Chanes, C.R.; Ribeiro, B.M.; Gilbertson, R.L.; Boiteux, L.S.; de Cássia Pereira-Carvalho, R. Nanopore sequencing of a novel bipartite New World begomovirus infecting cowpea. *Arch. Virol.* **2019**, *164*, 1907–1910, doi:10.1007/s00705-019-04254-5.
76. Leiva, A.M.; Siriwan, W.; Lopez-Alvarez, D.; Barrantes, I.; Hemniam, N.; Saokham, K.; Cuellar, W.J. Nanopore-Based Complete Genome Sequence of a Sri Lankan Cassava Mosaic Virus (Geminivirus) Strain from Thailand. *Microbiol. Resour. Announc.* **2020**, *9*, doi:10.1128/mra.01274-19.
77. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410, doi:10.1016/S0022-2836(05)80360-2.
78. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760, doi:10.1093/bioinformatics/btp324.
79. Finn, R.D.; Clements, J.; Eddy, S.R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **2011**, *39*, W29–W37, doi:10.1093/nar/gkr367.
80. Stobbe, A.H.; Daniels, J.; Espindola, A.S.; Verma, R.; Melcher, U.; Ochoa-Corona, F.; Garzon, C.; Fletcher, J.; Schneider, W. E-probe Diagnostic Nucleic acid Analysis (EDNA): A theoretical approach for handling of next generation sequencing data for diagnostics. *J. Microbiol. Methods* **2013**, *94*, 356–366, doi:10.1016/j.mimet.2013.07.002.
81. Punta, M.; Coggill, P.C.; Eberhardt, R.Y.; Mistry, J.; Tate, J.; Boursnell, C.; Pang, N.; Forslund, K.; Ceric, G.; Clements, J.; et al. The Pfam protein families database. *Nucleic Acids Res.* **2012**, *40*, 290–301, doi:10.1093/nar/gkr1065.
82. Marchler-Bauer, A.; Panchenko, A.R.; Shoemarker, B.A.; Thiessen, P.A.; Geer, L.Y.; Bryant, S.H. CDD: A database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.* **2002**, *30*, 281–283, doi:10.1093/nar/30.1.281.
83. Agranovsky, A.A.; Boyko, V.P.; Karasev, A. V.; Koonin, E. V.; Dolja, V. V. Putative 65 kDa protein of beet yellows closterovirus is a homologue of HSP70 heat shock proteins. *J. Mol. Biol.* **1991**, *217*, 603–610, doi:10.1016/0022-2836(91)90517-A.
84. Tangherlini, M.; Dell'Anno, A.; Zeigler Allen, L.; Riccioni, G.; Corinaldesi, C. Assessing viral taxonomic composition in benthic marine ecosystems: Reliability and efficiency of different bioinformatic tools for viral metagenomic analyses. *Sci. Rep.* **2016**, *6*, doi:10.1038/srep28428.
85. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and applications. *BMC Bioinformatics* **2009**, *10*, 1–9, doi:10.1186/1471-2105-10-421.
86. Kent, W.J. BLAT---The BLAST-Like Alignment Tool. *Genome Res.* **2002**, *12*, 656–664, doi:10.1101/gr.229202.
87. Buchfink, B.; Xie, C.; Huson, D.H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **2014**, *12*, 59–60, doi:10.1038/nmeth.3176.
88. Langmead, B.; Trapnell, C.; Pop, M.; Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the

human genome. *Genome Biol.* **2009**, *10*, R25, doi:10.1186/gb-2009-10-3-r25.

89. Mistry, J.; Finn, R.D.; Eddy, S.R.; Bateman, A.; Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **2013**, *41*, e121--e121, doi:10.1093/nar/gkt263.

90. Bzhalava, Z.; Hultin, E.; Dillner, J. Extension of the viral ecology in humans using viral profile hidden Markov models. *PLoS One* **2018**, *13*, e0190938, doi:10.1371/journal.pone.0190938.

91. Wood, D.E.; Salzberg, S.L. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **2014**, *15*, R46, doi:10.1186/gb-2014-15-3-r46.

92. Wood, D.E.; Lu, J.; Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **2019**, *20*, 1–13, doi:10.1186/s13059-019-1891-0.

93. Menzel, P.; Ng, K.L.; Krogh, A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* **2016**, *7*, doi:10.1038/ncomms11257.

94. Flygare, S.; Simmon, K.; Miller, C.; Qiao, Y.; Kennedy, B.; Di Sera, T.; Graf, E.H.; Tardif, K.D.; Kapusta, A.; Rynearson, S.; et al. Taxonomer: An interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. *Genome Biol.* **2016**, *17*, 1–18, doi:10.1186/s13059-016-0969-1.

95. Baizan-Edge, A.; Cock, P.; MacFarlane, S.; McGavin, W.; Torrance, L.; Jones, S. Kodoja: A workflow for virus detection in plants using k-mer analysis of RNA-sequencing data. *J. Gen. Virol.* **2019**, *100*, 533–542, doi:10.1099/jgv.0.001210.

96. Tampuu, A.; Bzhalava, Z.; Dillner, J.; Vicente, R. ViraMiner: Deep learning on raw DNA sequences for identifying viral genomes in human samples. *PLoS One* **2019**, *14*, 1–17, doi:10.1371/journal.pone.0222271.

97. Ren, J.; Song, K.; Deng, C.; Ahlgren, N.A.; Fuhrman, J.A.; Li, Y.; Xie, X.; Poplin, R.; Sun, F. Identifying viruses from metagenomic data using deep learning. *Quant. Biol.* **2020**, *8*, 64–77, doi:10.1007/s40484-019-0187-4.

98. Abdelkareem, A.O.; Khalil, M.I.; Elaraby, M.; Abbas, H.; Elbehery, A.H.A. VirNet: Deep attention model for viral reads identification. In Proceedings of the 2018 13th International Conference on Computer Engineering and Systems (ICCES); IEEE: Cairo, Egypt, 2018; pp. 623–626.

99. Ren, Y.; Xu, Y.; Lee, W.M.; Di Bisceglie, A.M.; Fan, X. In-depth serum virome analysis in patients with acute liver failure with indeterminate etiology. *Arch. Virol.* **2020**, *165*, 127–135, doi:10.1007/s00705-019-04466-9.

100. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **2018**, *34*, 3094–3100, doi:10.1093/bioinformatics/bty191.

101. Warwick-Dugdale, J.; Solonenko, N.; Moore, K.; Chittick, L.; Gregory, A.C.; Allen, M.J.; Sullivan, M.B.; Temperton, B. Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. *PeerJ* **2019**, *2019*, 1–28, doi:10.7717/peerj.6800.

102. Lefkowitz, E.J.; Dempsey, D.M.; Hendrickson, R.C.; Orton, R.J.; Siddell, S.G.; Smith, D.B. Virus taxonomy: The database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res.* **2018**, *46*, D708–D717, doi:10.1093/nar/gkx932.

103. Davison, A.J. Journal of general virology - Introduction to 'ICTV virus taxonomy profiles.' *J. Gen. Virol.* **2017**, *98*, 1, doi:10.1099/jgv.0.000686.

104. Bao, Y.; Chetvernin, V.; Tatusova, T. Improvements to pairwise sequence comparison (PASC): a genome-based web tool for virus classification. *Arch. Virol.* **2014**, *159*, 3293–3304, doi:10.1007/s00705-014-2197-x.

105. Gibbs, A.J.; Hajizadeh, M.; Ohshima, K.; Jones, R.A.C. The Potyviruses: An Evolutionary Synthesis Is Emerging. *Viruses* **2020**, *12*, 132, doi:10.3390/v12020132.

106. Jones, S.; Baizan-Edge, A.; MacFarlane, S.; Torrance, L. Viral diagnostics in plants using next generation sequencing: Computational analysis in practice. *Front. Plant Sci.* **2017**, *8*, doi:10.3389/fpls.2017.01770.

107. Blawid, R.; Silva, J.M.F.; Nagata, T. Discovering and sequencing new plant viral genomes by next-generation sequencing: description of a practical pipeline. *Ann. Appl. Biol.* **2017**, *170*, 301–314, doi:10.1111/aab.12345.

108. Roenhorst, J.W.; de Krom, C.; Fox, A.; Mehle, N.; Ravnikar, M.; Werkman, A.W. Ensuring validation in diagnostic testing is fit for purpose: a view from the plant virology laboratory. *EPPO Bull.* **2018**, *48*, 105–115, doi:10.1111/epp.12445.
109. Simmonds, P.; Adams, M.J.; Benk, M.; Breitbart, M.; Brister, J.R.; Carstens, E.B.; Davison, A.J.; Delwart, E.; Gorbalenya, A.E.; Harrach, B.; et al. Consensus statement: Virus taxonomy in the age of metagenomics. *Nat. Rev. Microbiol.* **2017**, *15*, 161–168, doi:10.1038/nrmicro.2016.177.
110. Rwahnih, M. Al; Daubert, S.; Úrbez-Torres, J.R.; Cordero, F.; Rowhani, A. Deep sequencing evidence from single grapevine plants reveals a virome dominated by mycoviruses. *Arch. Virol.* **2011**, *156*, 397–403, doi:10.1007/s00705-010-0869-8.
111. Marzano, S.Y.L.; Domier, L.L. Novel mycoviruses discovered from metatranscriptomics survey of soybean phyllosphere phytobiomes. *Virus Res.* **2016**, *213*, 332–342, doi:10.1016/j.virusres.2015.11.002.
112. Kreuze, J. siRNA Deep Sequencing and Assembly: Piecing Together Viral Infections. In *Detection and Diagnostics of Plant Pathogens*; Gullino, M.L., Bonants, P.J.M., Eds.; Springer Netherlands: Dordrecht, 2014; pp. 21–38 ISBN 978-94-017-9020-8.
113. Massart, S.; Candresse, T.; Gil, J.; Lacomme, C.; Predajna, L.; Ravnikar, M.; Reynard, J.S.; Rumbou, A.; Saldarelli, P.; Škoric, D.; et al. A framework for the evaluation of biosecurity, commercial, regulatory, and scientific impacts of plant viruses and viroids identified by NGS technologies. *Front. Microbiol.* **2017**, *8*, doi:10.3389/fmicb.2017.00045.
114. Kreuze, J.F.; Perez, A.; Gargurevich, M.G.; Cuellar, W.J. Badnaviruses of Sweet Potato: Symptomless Coinhabitants on a Global Scale. *Front. Plant Sci.* **2020**, *11*, 1–13, doi:10.3389/fpls.2020.00313.
115. Koloniuk, I.; Thekke-Veetil, T.; Reynard, J.S.; Pleško, I.M.; Přibylová, J.; Brodard, J.; Kellenberger, I.; Sarkisova, T.; Špak, J.; Lamovšek, J.; et al. Molecular characterization of divergent closterovirus isolates infecting Ribes species. *Viruses* **2018**, *10*, doi:10.3390/v10070369.
116. Sõmera, M.; Kvarnheden, A.; Desbiez, C.; Blystad, D.R.; Sooväli, P.; Kundu, J.K.; Gantsovski, M.; Nygren, J.; Lecoq, H.; Verdin, E.; et al. Sixty years after the first description: Genome sequence and biological characterization of European wheat striate mosaic virus infecting cereal crops. *Phytopathology* **2020**, *110*, 68–79, doi:10.1094/PHYTO-07-19-0258-FI.
117. Hammond, J.; Adams, I.; Fowkes, A.R.; McGreig, S.; Botermans, M.; van Oorspronk, J.J.A.; Westenberg, M.; Verbeek, M.; Dullemans, A.M.; Stijger, C.C.M.M.; et al. Sequence analysis of 43-year old samples of Plantago lanceolata show that Plantain virus X is synonymous with Actinidia virus X and is widely distributed. *Plant Pathol.* **2020**, 1–10, doi:10.1111/ppa.13310.
118. Tamisier, L.; Haegeman, A.; Foucart, Y.; Fouillien, N.; Rwahnih, M. Al; Buzkan, N.; Candresse, T.; Chiumenti, M.; Jonghe, K. De; Lefebvre, M.; et al. Semi-artificial datasets as a resource for validation of bioinformatics pipelines for plant virus detection. *Zenodo (preprint)* **2020**, doi:10.5281/zenodo.4273791.
119. Martin, D.P.; Murrell, B.; Golden, M.; Khoosal, A.; Muhire, B. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol.* **2015**, *1*, doi:10.1093/ve/vev003.
120. Lole, K.S.; Bollinger, R.C.; Paranjape, R.S.; Gadkari, D.; Kulkarni, S.S.; Novak, N.G.; Ingersoll, R.; Sheppard, H.W.; Ray, S.C. Full-Length Human Immunodeficiency Virus Type 1 Genomes from Subtype C-Infected Seroconverters in India, with Evidence of Intersubtype Recombination. *J. Virol.* **1999**, *73*, 152–160, doi:10.1128/jvi.73.1.152-160.1999.
121. Simmonds, P.; Midgley, S. Recombination in the Genesis and Evolution of Hepatitis B Virus Genotypes. *J. Virol.* **2005**, *79*, 15467–15476, doi:10.1128/jvi.79.24.15467-15476.2005.
122. Routh, A.; Johnson, J.E. Discovery of functional genomic motifs in viruses with ViReMa-a virus recombination mapper-for analysis of next-generation sequencing data. *Nucleic Acids Res.* **2014**, *42*, 1–10, doi:10.1093/nar/gkt916.
123. Xu, C.; Sun, X.; Taylor, A.; Jiao, C.; Xu, Y.; Cai, X.; Wang, X.; Ge, C.; Pan, G.; Wang, Q.; et al. Diversity, Distribution, and Evolution of Tomato Viruses in China Uncovered by Small RNA Sequencing. *J. Virol.* **2017**, *91*, 1–14, doi:10.1128/JVI.00173-17.
124. Bertran, A.; Ciuffo, M.; Margaria, P.; Rosa, C.; Resende, R.O.; Turina, M. Host-specific accumulation and temperature effects on the generation of dimeric viral RNA species derived from the S-RNA of members of the Tospovirus genus. *J. Gen. Virol.*

2016, *97*, 3051–3062, doi:10.1099/jgv.0.000598.

125. Saitou, N.; Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **1987**, *4*, 406–425, doi:10.1093/oxfordjournals.molbev.a040454.

126. Guindon, S.; Dufayard, J.F.; Lefort, V.; Anisimova, M.; Hordijk, W.; Gascuel, O. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* **2010**, *59*, 307–321, doi:10.1093/sysbio/syq010.

127. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **2014**, *30*, 1312–1313, doi:10.1093/bioinformatics/btu033.

128. Ronquist, F.; Teslenko, M.; Van Der Mark, P.; Ayres, D.L.; Darling, A.; Höhna, S.; Larget, B.; Liu, L.; Suchard, M.A.; Huelsenbeck, J.P. Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **2012**, *61*, 539–542, doi:10.1093/sysbio/sys029.

129. Huson, D.H.; Beier, S.; Flade, I.; Górska, A.; El-Hadidi, M.; Mitra, S.; Ruscheweyh, H.J.; Tappu, R. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Comput. Biol.* **2016**, *12*, 1–12, doi:10.1371/journal.pcbi.1004957.

130. Suchard, M.A.; Lemey, P.; Baele, G.; Ayres, D.L.; Drummond, A.J.; Rambaut, A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **2018**, *4*, doi:10.1093/ve/vey016.

131. Hardy, O.J.; Vekemans, X. SPAGeDI: A versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol. Ecol. Notes* **2002**, *2*, 618–620, doi:10.1046/j.1471-8286.2002.00305.x.

132. Zheng, Y.; Gao, S.; Padmanabhan, C.; Li, R.; Galvez, M.; Gutierrez, D.; Fuentes, S.; Ling, K.S.; Kreuze, J.; Fei, Z. VirusDetect: An automated pipeline for efficient virus discovery using deep sequencing of small RNAs. *Virology* **2017**, *500*, 130–138, doi:10.1016/j.virol.2016.10.017.

133. Lefebvre, M.; Theil, S.; Ma, Y.; Candresse, T. The virannot pipeline: A resource for automated viral diversity estimation and operational taxonomy units assignation for virome sequencing data. *Phytobiomes J.* **2019**, *3*, 256–259, doi:10.1094/PBIOMES-07-19-0037-A.

134. Ho, T.; Tzanetakis, I.E. Development of a virus detection and discovery pipeline using next generation sequencing. *Virology* **2014**, *471–473*, 54–60, doi:10.1016/j.virol.2014.09.019.

135. Visser, M.; Burger, J.T.; Maree, H.J. Targeted virus detection in next-generation sequencing data using an automated e-probe based approach. *Virology* **2016**, *495*, 122–128, doi:10.1016/j.virol.2016.05.008.

136. Afgan, E.; Baker, D.; Batut, B.; Van Den Beek, M.; Bouvier, D.; Ech, M.; Chilton, J.; Clements, D.; Coraor, N.; Grüning, B.A.; et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* **2018**, *46*, W537–W544, doi:10.1093/nar/gky379.

137. Kalantar, K.L.; Carvalho, T.; de Bourcy, C.F.A.; Dimitrov, B.; Dingle, G.; Egger, R.; Han, J.; Holmes, O.B.; Juan, Y.F.; King, R.; et al. IDseq-An open source cloud-based pipeline and analysis service for metagenomic pathogen detection and monitoring. *Gigascience* **2020**, *9*, 1–14, doi:10.1093/gigascience/giaa111.